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

Background: LTI Systems and MJLS

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

$$x^k = (H)x^0 + \sum_{i=0}^{k-1} (H)^{k-i-1} Gu^i.$$  

• Let $z^k$ be a Markov chain sampled from a finite state space $S$. A MJLS is governed by the following state-space model: $z^{k+1} = H(z^k)x^k + Gu^k$ 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 $q$ and covariance $Q$ are available, where

$$q^k = E(z^k I_{(z^k = i)}) , \quad Q^k = E((z^k - q^k)(z^k - q^k)^T), \quad \mu^k = Ez^k,$$

$$Q^k = E((z^k - q^k)^T(z^k - q^k)) = [q_1^T \ldots q_{|S|}^T] = [q_1^T \ldots q_{|S|}^T].$$

TD learning under IID Assumption

Theorem 1 Consider a MJLS with $H_1 = I + \alpha A, G_1 = \alpha b$, and $y^k = 1$. Suppose $\{z^k\}$ is sampled from $N$ using an IID distribution $P(z^k = i) = p_i$. In addition, assume $\sum_i p_i = 0$. Then set $H_{21} = \alpha I + \alpha A$ and $H_{22} = I + \alpha (I + A)$, and

$$H = \alpha I + \alpha A, \quad I + \alpha (I + A) = \alpha I + \alpha A.$$ 

The input for the exact LTI model does not change with $k$. Therefore, if $\sigma(H_2)$ is Hurwitz, there exists sufficiently small $\epsilon > 0$ such that $\sigma(H_2) < \epsilon$. Hence as long as $\bar{A}$ is Hurwitz, there exists sufficiently small $\alpha$ such that $\sigma(H_2) < \epsilon$.

Stability Condition: LTI system (3) is stable if and only if $H_{22}$ is Hurwitz. For TD learning to converge, it is important to choose $\delta$ such that $\sigma(H_2) < 1$ for some given $\{A_i\}, \{b_i\}$ and $\{\epsilon_i\}$. Assuming $\alpha$ to be small, eigenvalue perturbation analysis to $H_2$ suggests: $\sigma(H_2) \approx 1 + \epsilon_2 (\epsilon_2 + O(\epsilon^2)).$ Therefore, as long as $\bar{A}$ is Hurwitz, there exists sufficiently small $\alpha$ such that $\sigma(H_2) < 1$.

Corollary 1 Consider TD update (2) with $A$ being Hurwitz. Suppose $\sigma(H_2) < 1$ and $P(z^k = i) = p_i$. Then $\delta^* = \lim_{k \to \infty} E(\|z^k - \theta^k\|^2)$ exists and is given by $\delta^* = \text{trace}(Q^\infty)$. Additionally, the following Mean Square TD error bound holds for some constant $C_0$ and any arbitrary small $\epsilon > 0$ (the rate $\sigma(H)$ is precise):

$$\delta^* - C_0 (\sigma(H) + \epsilon)^2 \leq E(\|z^k - \theta^k\|^2) \leq \delta^* + C_0 (\sigma(H) + \epsilon)^2$$

Key Trade-off: For small $\alpha$, one can use perturbation to show $\lim_{k \to \infty} E(\|z^k - \theta^k\|^2) = O(\epsilon)$ and $\sigma(H) = 1 + \epsilon_2(\epsilon_2 + O(\epsilon^2)).$ This gives the standard error rate for TD learning.