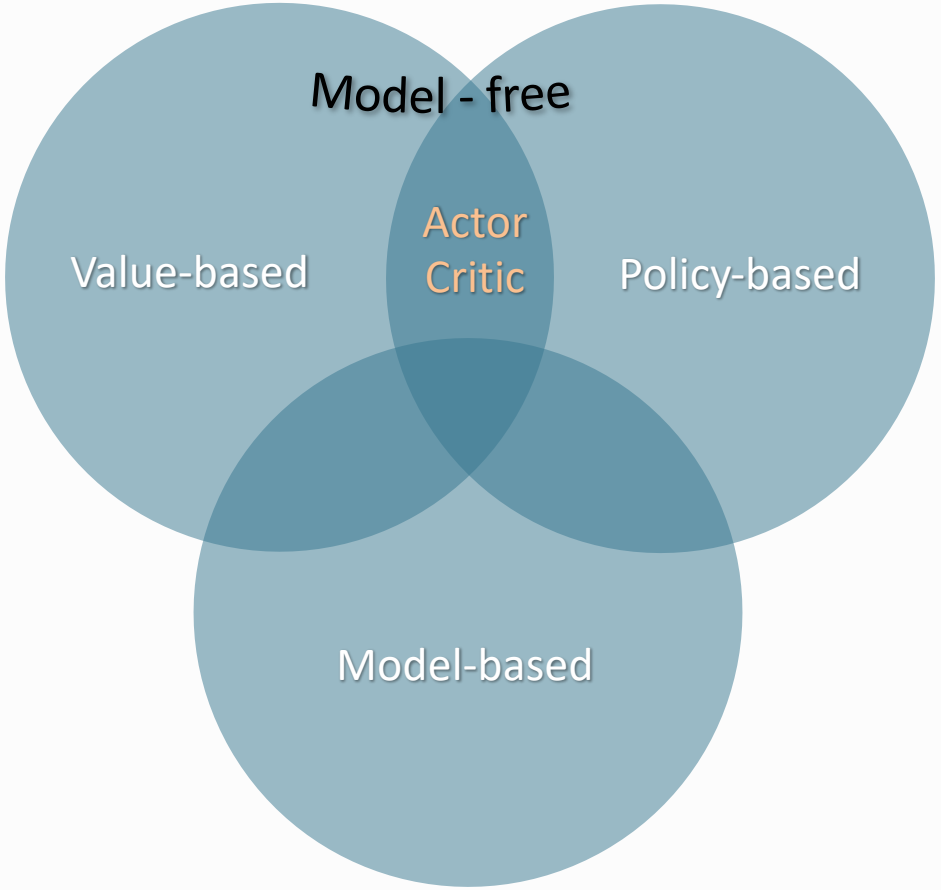
EPFL Summer School on Data Science, Optimization and Operations Research
August 15-20, 2021

Lecture 2: RL from Control and Dynamical Systems' Perspective

Niao He, D-INFK, ETH Zurich

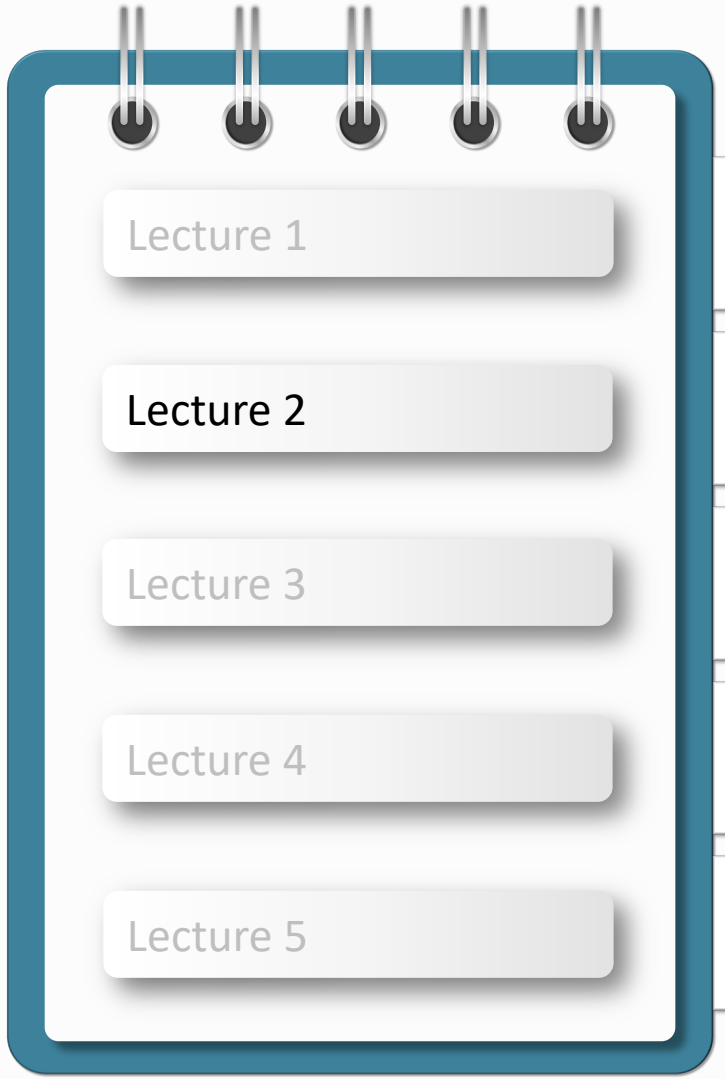
Recap: Reinforcement Learning Approaches



- **Value-based RL**
  - Estimate the optimal value function $Q^*(s, a)$
  - Example: Q-learning
- **Policy-based RL**
  - Search directly the optimal policy $\pi^*(\cdot\,|s)$
  - Example: Policy Gradient Method
- **Model-based RL**
  - First estimate the model P, R and then do planning

What are the convergence behaviors of these algorithms?

Outline of Lecture Series



Introduction to RL

RL from Control Perspectives
- *Value-based RL*

RL from Optimization Perspectives
- *Policy-based RL*

RL from Learning Perspectives
- *Deep RL*

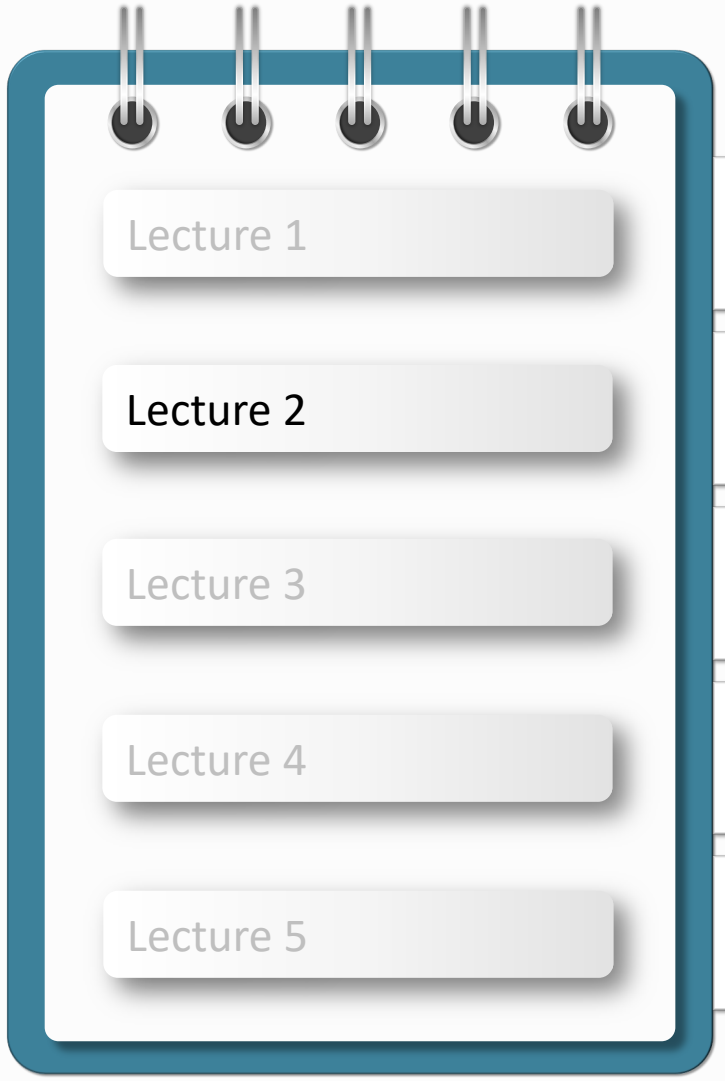
RL from Game Perspectives

Focus:
Unified control-theoretic analysis of model-free value-based methods

- TD learning
- Q learning
- Double Q learning
- Variants w/o function approximation

Challenge:
- ***Not*** *stochastic gradient methods* (stochastic semi-gradient)
- ***Not*** *i.i.d. sampling* (Markovian data)
- *Varying dynamics* (linear or nonlinear)

Outline of Lecture Series



Lecture 1

Lecture 2

Lecture 3

Lecture 4

Lecture 5

Introduction to RL

RL from Control Perspectives
- *Value-based RL*

RL from Optimization Perspectives
- *Policy-based RL*

RL from Learning Perspectives
- *Deep RL*

RL from Game Perspectives

Focus:

Unified control-theoretic convergence analysis of model-free value-based methods

- TD learning
- Q learning
- Double Q learning
- Variants w/o function approximation

Stochastic Approximation algorithms:

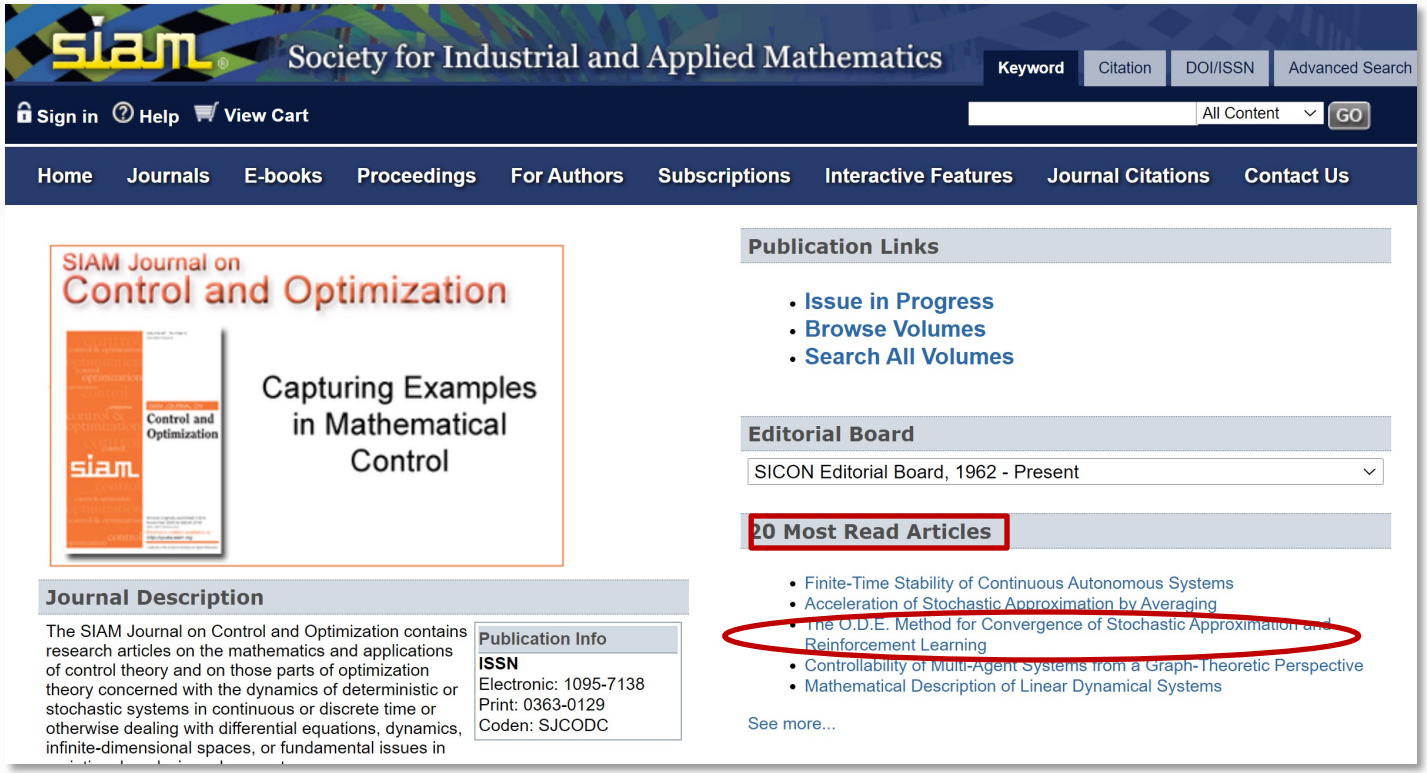
$$x_{k+1} = x_k + \alpha_k(h(x_k) + \epsilon_{k+1}), k = 0, 1, \ldots$$

Notation Recap

- **MDP** $(S, \mathcal{A}, P, R, \mu, \gamma)$

State value function:	$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \vert s_0 = s \right]$
State-action value function:	$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \vert s_0 = s, a_0 = a \right]$
Optimal value function:	$V^*(s) := \max_\pi V^\pi(s), \quad Q^*(s, a) := \max_\pi Q^\pi(s, a)$
Optimal policy:	$\pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q^*(s, a)$
Bellman equation:	$V^\pi(s) = \mathbb{E}_{a \sim \pi(s)} \left[R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot \vert s, a)} V^\pi(s') \right]$
Bellman optimality:	$Q^*(s, a) = R(s, a) + \mathbb{E}_{s' \vert s, a} \left[\gamma \max_{a' \in \mathcal{A}} Q^*(s', a') \right]$

The ODE Method



SIAM J. CONTROL OPTIM.
Vol. 38, No. 2, pp. 447–469

THE O.D.E. METHOD FOR CONVERGENCE OF STOCHASTIC APPROXIMATION AND REINFORCEMENT LEARNING*

V. S. BORKAR† AND S. P. MEYN‡

Abstract. It is shown here that stability of the stochastic approximation algorithm is implied by the asymptotic stability of the origin for an associated ODE. This in turn implies convergence of the algorithm. Several specific classes of algorithms are considered as applications. It is found that the results provide (i) a simpler derivation of known results for reinforcement learning algorithms; (ii) a proof for the first time that a class of asynchronous stochastic approximation algorithms are convergent without using any a priori assumption of stability; (iii) a proof for the first time that asynchronous adaptive critic and Q-learning algorithms are convergent for the average cost optimal control problem.

Key words. stochastic approximation, ODE method, stability, asynchronous algorithms, reinforcement learning

AMS subject classifications. 62L20, 93E25, 93E15

PII. S0363012997331639

1. Introduction. The stochastic approximation algorithm considered in this paper is described by the d-dimensional recursion

$$X(n+1) = X(n) + a(n)\big[h\big(X(n)\big) + M(n+1)\big], \quad n \ge 0, \tag{1.1}$$

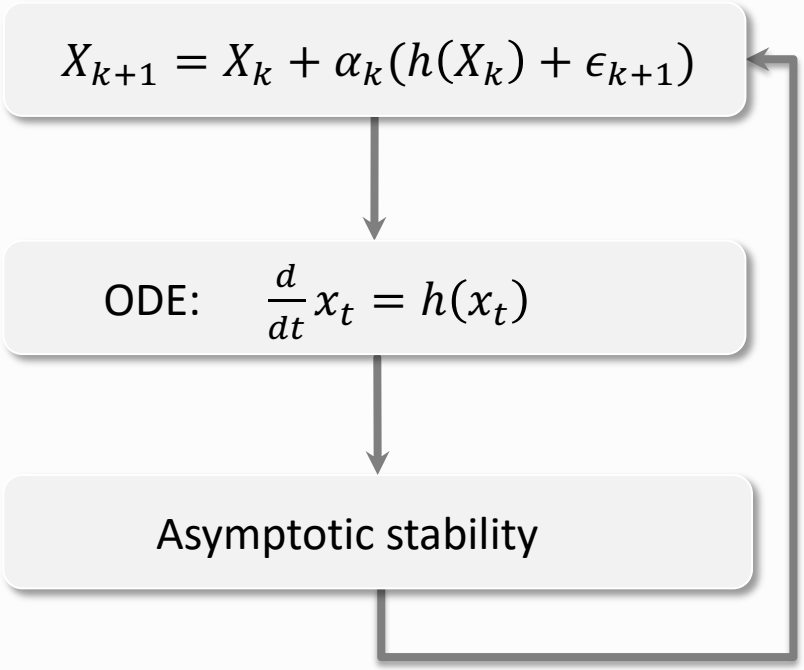
where $X(n) = [X_1(n), \ldots, X_d(n)]^T \in \mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}^d$, and $\{a(n)\}$ is a sequence of positive numbers. The sequence $\{M(n) : n \ge 0\}$ is uncorrelated with zero mean.

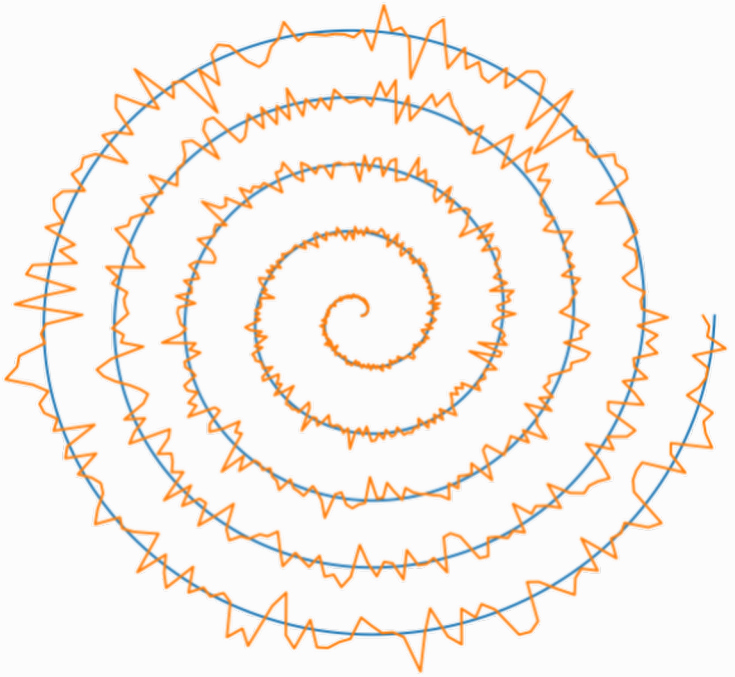
Though more than four decades old, the stochastic approximation algorithm is now of renewed interest due to novel applications to reinforcement learning [20] and as a model of learning by boundedly rational economic agents [19]. Traditional convergence analysis usually shows that the recursion (1.1) will have the desired asymptotic behavior provided that the iterates remain bounded with probability one, or that they visit a prescribed bounded set infinitely often with probability one [3, 14]. Under such stability or recurrence conditions one can then approximate the sequence $\boldsymbol{X} = \{X(n) : n \ge 0\}$ with the solution to the ordinary differential equation (ODE)

$$\dot{x}(t) = h\big(x(t)\big) \tag{1.2}$$

with identical initial conditions $x(0) = X(0)$.

The ODE Method: Key Idea





Dynamical system, $\frac{d}{dt}x_t = h(x_t)$, is *globally asymptotically stable* if $x_t \to x^*$ for any x_0.

The ODE Method: Borkar and Meyn Theorem

SA: $X_{k+1} = X_k + \alpha_k(h(X_k) + \epsilon_{k+1})$

[Borkar and Meyn Theorem, 2000]

Under the following conditions:

a) Global Lipschitz continuity of the mapping h

b) Robbins-Monro stepsize: $\sum \alpha_k = \infty, \sum \alpha_k^2 < \infty$

c) Bounded noise of martingale difference: $E[\|\epsilon_{k+1}\|^2 | G_k] \leq C_0(1 + \|X_k\|^2), \forall k \geq 0$

d) *Asymptotic stability of the limiting ODE*: $\dot{x}_t = h_\infty(x_t) \coloneqq \lim_{c \to \infty} \frac{h(cx)}{c}$

e) *Global asymptotic stability of the original ODE*: $\dot{x}_t = h(x_t)$

we have $X_k \to x^*$ as $k \to \infty$.

Stability of Linear Systems

- **Linear System:**

$$\frac{d}{dt}x_t = Ax_t$$

The origin is an asymptotically stable equilibrium point if and only if A is *Hurwitz*.
Or equivalently, there exists $M = M^T \succ 0$ such that $A^T M + MA \prec 0$.

- A matrix is Hurwitz if all eigenvalues have strictly negative real parts.
- Lyapunov function: $V(x) = x^T Mx, \quad \frac{dV(x_t)}{dt} = x_t^T (A^T M + MA)x_t < 0, \; V(x_t) \to 0.$
- Applications to TD-learning variants: TD(0), TD(λ), GTD, TDC, A-TD, D-TD, etc.

Convergence of TD-learning

- TD-learning with Linear Function Approximation [Tsitsiklis & Van Roy, 1997]

$$\theta_{k+1} = \theta_k + \alpha_k \phi(s_k)[r(s_k, a_k) + \gamma \phi(s_{k+1})^T \theta_k - \phi(s_k)^T \theta_k]$$

$$\frac{d}{dt}(\theta_t - \theta^*) = A(\theta_t - \theta^*)$$

$A = \Phi^T D(\gamma P^\pi - I)\Phi$ is *Hurwitz* if Φ is full rank.

- $\Phi^T = [\phi(1), \phi(2), \dots, \phi(|S|)]$
- $P^\pi(s, s') = \sum_a P(s'|s, a)\pi(a|s)$
- $D = diag(d), d = dP^\pi$, state stationary distribution

Convergence of Double TD-learning

- Double TD-learning with Linear Function Approximation [Lee & He, 2019]

$$\theta_{k+1}^A = \theta_k^A + \alpha_k \phi(s_k)\left[r(s_k,a_k) + \gamma\phi(s_{k+1})^T\theta_k^B - \phi(s_k)^T\theta_k^A\right] + \delta(\theta_k^B - \theta_k^A)$$
$$\theta_{k+1}^B = \theta_k^B + \alpha_k \phi(s_k)\left[r(s_k,a_k) + \gamma\phi(s_{k+1})^T\theta_k^A - \phi(s_k)^T\theta_k^B\right] + \delta(\theta_k^A - \theta_k^B)$$

$$\frac{d}{dt}(\theta_t - \theta^*) = A(\theta_t - \theta^*), \quad A = B + C^T BC, with\ B = \begin{bmatrix} -\Phi^T D\Phi & \gamma\Phi^T DP^\pi\Phi \\ \delta I & -\delta I \end{bmatrix}, C = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$$

A is *Hurwitz* if Φ is full rank, $\delta > 0$.

Theorem. Under Robbins-Monro stepsize and assume the Markov chain under policy π is ergodic, we have

$$\theta_k \to \theta^* \text{ almost surely, as } k \to \infty,$$

Here θ^* is the solution of the projected Bellman equation: $\Phi\theta = \Pi(R^\pi + \gamma P^\pi \Phi\theta)$.

Stability of Nonlinear Systems

- **Nonlinear System:**

$$\frac{d}{dt}x_t = h(x_t)$$

[Khalil, 2002]

The origin is unique, globally asymptotically stable if there exists a twice differentiable Lyapunov function $V(x)$ such that

$$k_1\|x\|^\alpha \le V(x) \le k_2\|x\|^\alpha$$
$$\frac{dV}{dx}h(x) \le -k_3\|x\|^\alpha$$

for some positive constants α, k_1, k_2, k_3.

- Sufficiency but not necessity.
- Application: tabular Q-learning [Borkar & Meyn, 2000]
- Application: Q-learning with linear function approximation [Melo et al., 2008] [Wang & Giannakis, 2020]

Stability of Linear Switching Systems

- **Linear switching system**:

$$\frac{d}{dt}x_t = A_{\sigma_t}x_t$$

  - coupling between continuous dynamics and discrete events (switching)
  - σ_t: switching signal $\in \{1,2,\cdots,M\}$; $\{A_1,\dots,A_M\}$ subsystem matrices
  - $\sigma_t = \sigma(x_t)$: state-feedback switching signal

[Lin and Antsaklis, 2009]

The origin is the unique globally asymptotically stable equilibrium point if and only if there exists a full column rank matrix L and a family of NRD matrices $\{\bar{A}_1,\dots,\bar{A}_M\}$ such that

$$LA_\sigma = \bar{A}_\sigma L, \forall \sigma \in \{1,2,\dots,M\}.$$

Negative Row Dominant Diagonal (NRD) matrix A: $a_{ii} + \sum_{j\neq i}|a_{ij}| < 0, \forall i.$

Switching System Model of Q-Learning

- **Q-learning**

$$Q_{k+1}(s_k, a_k) = Q_k(s_k, a_k) + \alpha_k(r(s_k, a_k) + \gamma \max_{a'} Q_k(s_{k+1}, a') - Q_k(s_k, a_k))$$

- **Dynamical system**

$$\frac{d}{dt}(Q_t - Q^*) = \left(\gamma DP\Pi_{\pi_{Q_t}} - D\right)(Q_t - Q^*) + \gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi^*})Q^*$$

  - Greedy policy: $\pi_{Q_t}(s) = \text{argmax}_a Q_t(s, a)$
  - Diagonal elements of D : state-action distribution
  - $P = \begin{bmatrix} \vdots \\ P_a \\ \vdots \end{bmatrix}$, P_a=transition probability matrix for taking action a
  - $\Pi_\pi \coloneqq [\cdots \quad \Gamma_a \quad \cdots], [\Gamma_a]_{(s,a')} = 1 \; if \;\; \pi(s) = a'$ and 0 otherwise

Switching System Model of Q-Learning

- **Q-learning**

$$Q_{k+1}(s_k, a_k) = Q_k(s_k, a_k) + \alpha_k(r(s_k, a_k) + \gamma \max_{a'} Q_k(s_{k+1}, a') - Q_k(s_k, a_k))$$

- **Dynamical system**

$$\frac{d}{dt}(Q_t - Q^*) = \left(\gamma DP\Pi_{\pi_{Q_t}} - D\right)(Q_t - Q^*) + \gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi^*})Q^*$$

Affine switching system

$$\frac{d}{dt}x_t = A_{\sigma(x_t)}x_t + b_{\sigma(x_t)}$$

  - $x_t = Q_t - Q^*, \sigma(x_t) = \psi(\pi_{Q_t}), \pi_{Q_t}(s) = \text{argmax}_a Q_t(s, a)$
  - ψ: deterministic policy → integer

Stability Analysis: Upper and Lower Comparison Systems

- **Upper comparison system (linear switching system)**

$$\frac{d}{dt}(Q_t - Q^*) = \left(\gamma DP\Pi_{\pi_{Q_t}} - D\right)(Q_t - Q^*)$$

$$\geq$$

- **Original affine switching system**

$$\frac{d}{dt}(Q_t - Q^*) = \left(\gamma DP\Pi_{\pi_{Q_t}} - D\right)(Q_t - Q^*) + \gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi^*})Q^*$$

$$\geq$$

- **Lower comparison system (linear system)**

$$\frac{d}{dt}(Q_t - Q^*) = \left(\gamma DP\Pi_{\pi_{Q^*}} - D\right)(Q_t - Q^*)$$

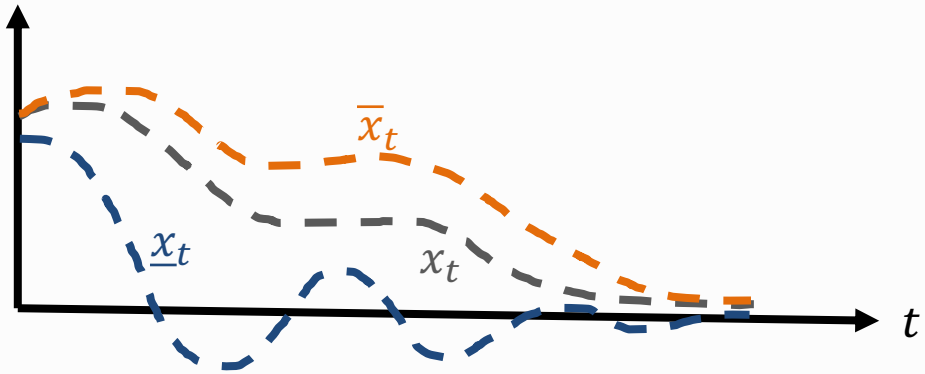
Note $A_\sigma = \gamma D(P\Pi_\sigma - I)$ $is\ NRD: [A_\sigma]_{ii} + \sum_{j \neq i}\left|[A_\sigma]_{ij}\right| \leq \gamma - 1 < 0, \forall\sigma.$

Stability Analysis: Vector Comparison Principle

$$\underbrace{\frac{d}{dt}\underline{x}_t = \underline{f}(\underline{x}_t)}_{\text{Linear system } A \text{ is Hurwitz}} \leq \frac{d}{dt}x_t = f(x_t) \leq \underbrace{\frac{d}{dt}\overline{x}_t = \overline{f}(\overline{x}_t)}_{\text{Linear switching system } A_\sigma \text{ is } NRD}$$

$$\downarrow t \to \infty \qquad\qquad \downarrow t \to \infty$$

$$\underline{x}_t \to 0 \qquad x_t \to 0 \qquad \overline{x}_t \to 0$$



[Vector Comparison Principle]

If $\bar{f}$ and f are globally Lipschitz continuous, $\bar{f}$ is quasi-monotone increasing, then

$$f \leq \bar{f}, x_0 \leq \bar{x}_0 \Rightarrow x_t \leq \bar{x}_t, \forall t \geq 0.$$

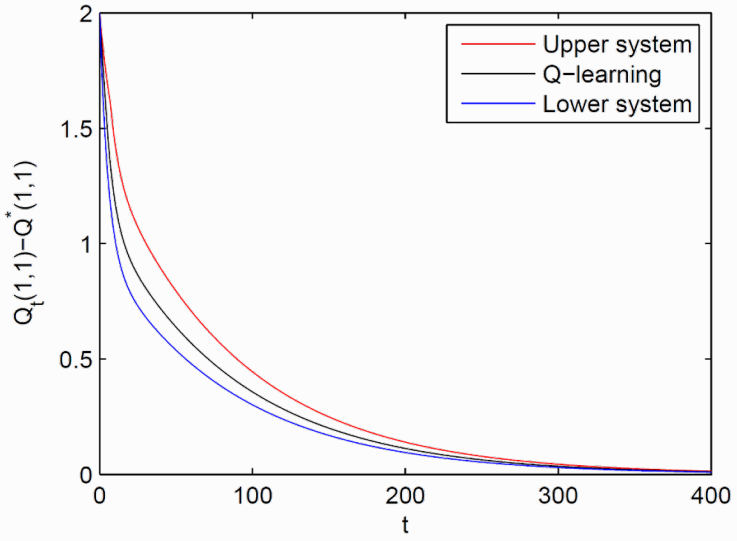
Asymptotic Stability and Convergence

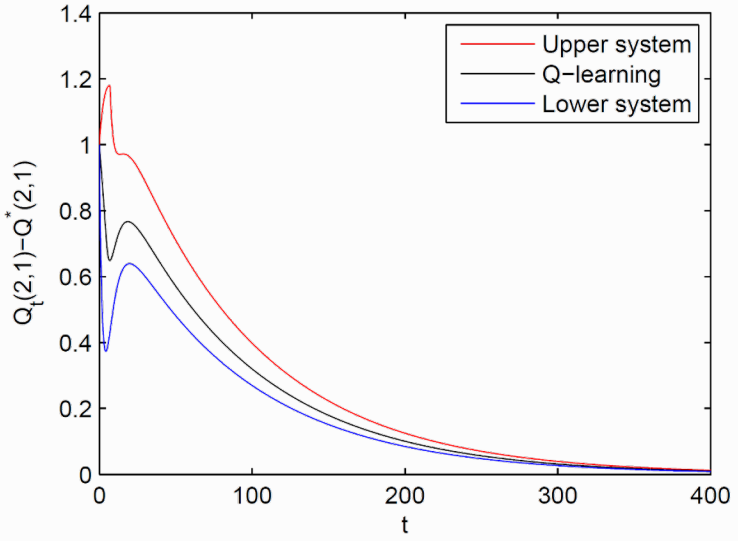
Theorem. Under Robbins-Monro stepsize and ergodicity assumption, $Q_k \to Q^*$ almost surely, as $k \to \infty$.

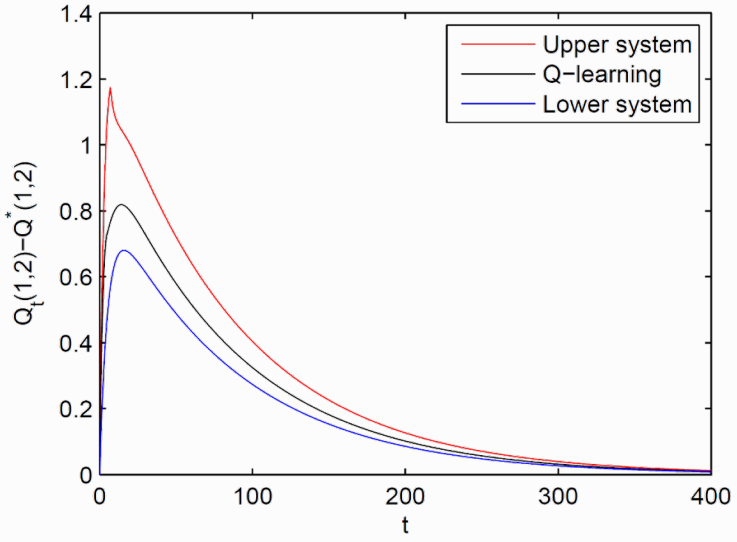
- Other proofs
  - The original proof [Watkins and Dayan, 1992]
  - Stochastic-approximation-based approach [Jaakkola et al., 1994] [Tsitsiklis, 1994]
  - Finite-time analysis: [Szepesvári, 1998][Even-Dar & Mansour, 2003]
  - Recent work: [Qu & Wierman, 2020] [Li et al., 2020]

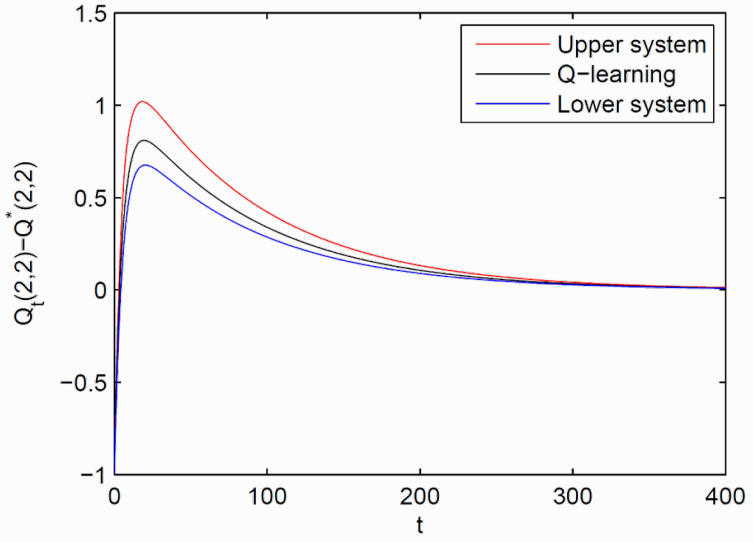
- Extensions:
  - Target-based Q-learning algorithms
  - Q-learning with linear function approximation

Illustration









Q-learning with Linear Function Approximation

- **Algorithm:**

$$\theta_{k+1} = \theta_k + \alpha_k \phi(s_k, a_k)[r(s_k, a_k) + \gamma \max_{a'}(\Phi\theta_k)(s_{k+1}, a') - (\Phi\theta_k)(s_k, a_k)]$$

- **Switching system:**

$$\frac{d}{dt}\theta_t = \left(\gamma \Phi^T D P \Pi_{\pi(\theta_t)} \Phi - \Phi^T D \Phi\right)\theta_t + \Phi^T D R$$

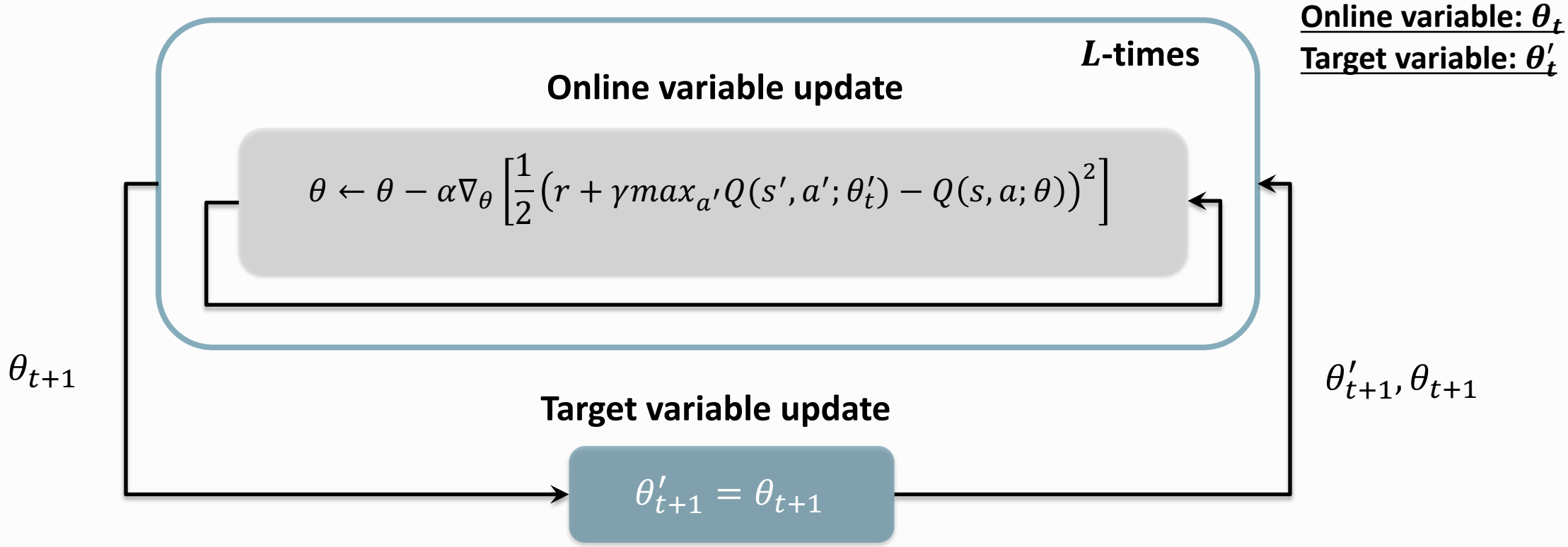
- **New sufficient condition:**

$$-\phi_i^T D \phi_i + \gamma \phi_i^T D P \Pi_\pi \sum \phi_j < 0, \forall \text{ admissible } \pi$$

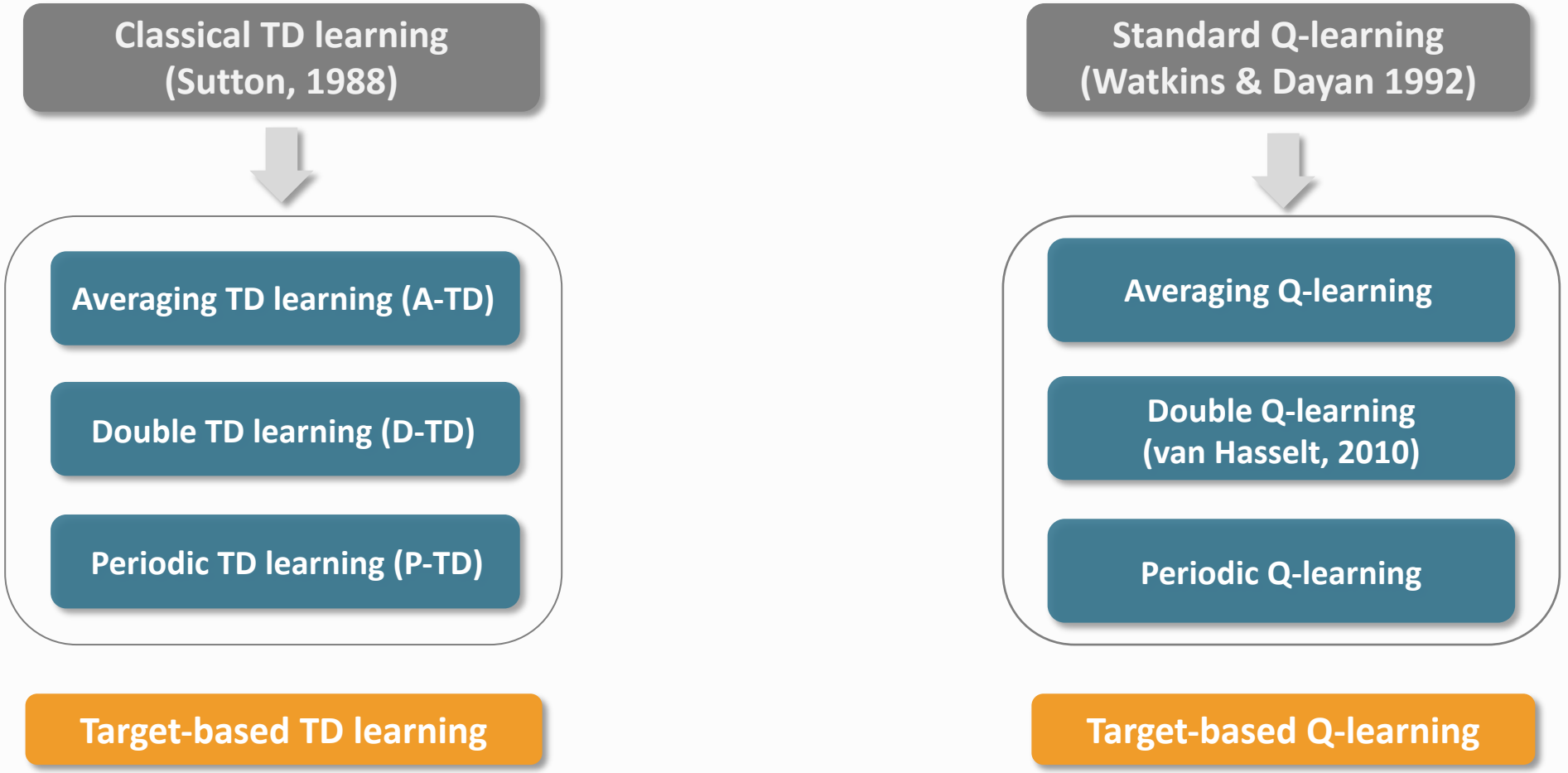
  - Under the above condition, we can easily show that for $\theta_k \to \theta^*$ as $k \to \infty$.
  - Less conservative than the Melo's condition.

Deep Q-learning

- The use of target network is pervasive in DQN-like algorithms.

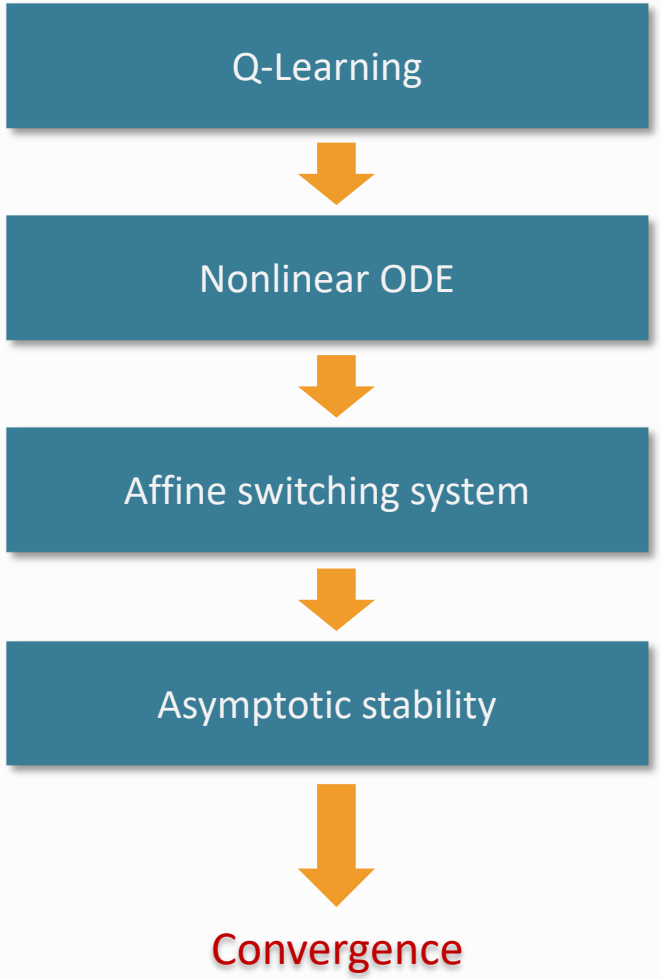


Target-based Q-learning



Lee and **H.**, "Target-based temporal difference learning," ICML, 2019.
Lee and **H.**, "Periodic Q-learning," L4DC, 2020.

The Roadmap



$$Q_{k+1}(s,a) = Q_k(s,a) - \alpha_k\left(r + \gamma \max_{a'} Q_k(s',a') - Q_k(s,a)\right)$$

$$\frac{d}{dt}(Q_t - Q^*) = \left(\gamma DP\Pi_{\pi_{Q_t}} - D\right)(Q_t - Q^*) + \gamma DP(\Pi_{\pi_{Q_t}} - \Pi_{\pi^*})Q^*$$

$$\frac{d}{dt}x_t = A_{\sigma(x_t)}x_t + b_{\sigma(x_t)}$$

$$x_t \to x^*$$

Dynamical Systems Perspective for RL

RL algorithms

$$x_{k+1} = x_k + \alpha_k H(x_k, s_k, a_k, s_{k+1})$$

- TD learning (tabular, LFA)
- Q learning (tabular, LFA)
- RL with nonlinear FA

Dynamical Systems (continuous/discrete)

$$\frac{d}{dt} x_t = h(x_t)$$

- Time-invariant Linear System
  Markov Jump Linear System
- Switched Affine System
- Nonlinear System

Dynamical Systems Perspective for RL

- **Advantages:**
  - Systematic and unified analysis
  - Rich control theory and tools
  - No need for objective/gradients/regularizations
  - Characterization of exact behavior
  - Tight conditions and weak assumptions

- **Challenges:**
  - Stability of nonlinear systems
  - Characterization of non-asymptotic behaviors

From Asymptotic Convergence to Asymptotic Covariance

- Stochastic approximation algorithm:

$$\theta_{k+1} = \theta_k + \alpha_k h(\theta_k) + \epsilon_{k+1}$$

- Central Limit Theorem:

$$\theta_k \to \theta^*, \text{almostly surely}$$

$$\sqrt{k}(\theta_k - \theta^*) \to N(0, \Sigma) \text{ in distribution}$$

Asymptotic Variance and Lyapunov Equation

- Consider the linear stochastic approximation

$$\theta_{k+1} = \theta_k + \frac{g}{k}(A(Y_k)\theta_k + b(Y_k))$$

  - Assume $\theta_k \to \theta^* = 0$
  - Y_k is an irreducible aperiodic Markov chain, $A = E[A(Y_\infty)], \Sigma_b = \sum_{k=2}^{\infty} E[b(Y_k)b(Y_1)^T]$
  - Define the asymptotic covariance $\Sigma_\infty := \lim_{k\to\infty} n\, E[\theta_k \theta_k^T]$

[Kushner and Yin, 2003; Chen et al., 2020]

If $\frac{1}{2}I + g\,A$ is Hurwitz, then Σ_∞ is the unique solution to the Lyapunov equation:

$$\left(\frac{1}{2}I + g\,A\right)\Sigma_\infty + \Sigma_\infty\left(\frac{1}{2}I + g\,A^T\right) + g^2\Sigma_b = 0$$

Q-learning vs. Double Q-learning

- **Q-learning with LFA:**

$$\theta_{k+1} = \theta_k + \alpha_k \phi(s_k, a_k)[r(s_k, a_k) + \gamma H(\theta_k, \theta_k, s_{k+1}) - \phi(s_k, a_k)^T \theta_k]$$

- **Double Q-learning with LFA:**

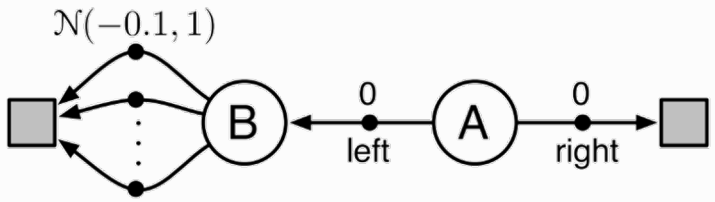
$$\theta^A_{k+1} = \theta^A_k + \beta_k \delta_k \phi(s_k, a_k)[r(s_k, a_k) + \gamma H(\theta^A_k, \theta^B_k, s_{k+1}) - \phi(s_k, a_k)^T \theta^A_k$$

$$\theta^B_{k+1} = \theta^B_k + (1 - \beta_k)\delta_k \phi(s_k, a_k)[r(s_k, a_k) + \gamma H(\theta^B_k, \theta^A_k, s_{k+1}) - \phi(s_k, a_k)^T \theta^B_k$$

$$H(\theta_1, \theta_2, s) = \phi\left(s, \arg\max_a \phi(s, a)^T \theta_1\right)^T \theta_2, \quad \beta_k \sim \text{Bernoulli}\left(\frac{1}{2}\right) \text{ i.i.d.}$$

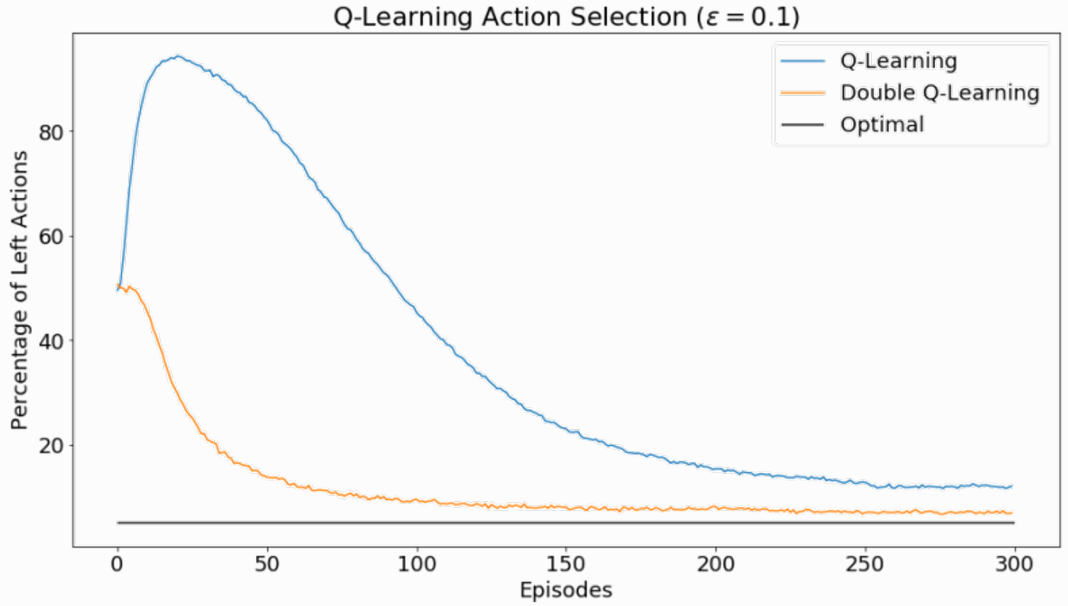
Q-learning vs. Double Q-learning

- Folklore: double Q-learning helps reduce the maximization bias.



Example

[Sutton and Barto, 2018]



The Asymptotic Mean-Square Errors of Q-learning Algorithms

Theorem Set $\alpha_k = \frac{g}{k}, \delta_k = \frac{2g}{k}$ and assume both Q-learning and Double Q-learning converge. Under mild conditions, we have

$$AMSE(\theta^A) = AMSE(\theta^B) \geq AMSE(\theta) + c_0 g$$

$$AMSE\left(\frac{\theta^A + \theta^B}{2}\right) = \mathrm{AMSE}(\theta)$$

- Q-learning: $AMSE(\theta) \coloneqq \lim_{k\to\infty} kE\|\theta_k - \theta^*\|^2$
- Double Q-learning: $AMSE(\theta^A) \coloneqq \lim_{k\to\infty} kE\left\|\theta_k^A - \theta^*\right\|^2$
- Double Q-learning: $AMSE(\theta^B) \coloneqq \lim_{k\to\infty} kE\left\|\theta_k^B - \theta^*\right\|^2$
- Double Q-learning with average estimator: $AMSE\left(\frac{\theta^A+\theta^B}{2}\right) \coloneqq \lim_{k\to\infty} kE\left\|\frac{\theta_k^A+\theta_k^B}{2} - \theta^*\right\|^2$

Proof Sketch

- Lyapunov equation for Q learning

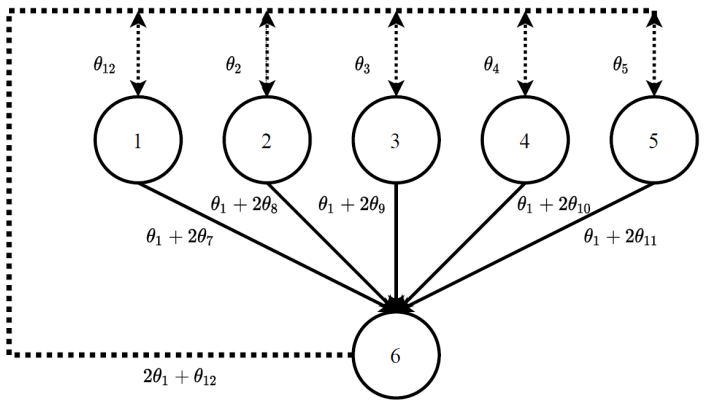
$$\left(\frac{1}{2}I + g\,A\right)\Sigma_\infty + \Sigma_\infty\left(\frac{1}{2}I + g\,A^T\right) + g^2\Sigma_b = 0, \qquad A = \Phi D(\gamma P\Pi_{\pi*} - I)\Phi^{\mathrm{T}}$$

- Lyapunov equation for Double Q learning:

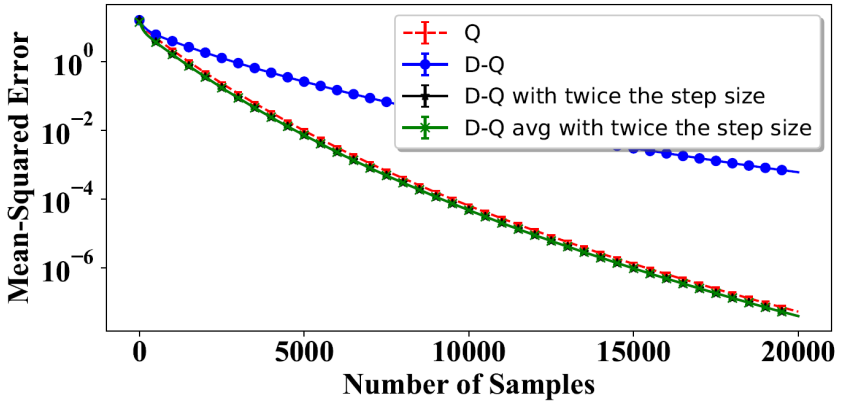
$$\left(\frac{1}{2}I + g\,A_D\right)\Sigma_\infty^D + \Sigma_\infty^D\left(\frac{1}{2}I + g\,A_D^T\right) + g^2\Sigma_b^D = 0, \quad A_D = \begin{bmatrix} -\Phi^T D\Phi & \gamma\Phi DP\Pi_{\pi^*}\Phi^T \\ \Phi DP\Pi_{\pi^*}\Phi^T & -\Phi^T D\Phi \end{bmatrix}$$

- Key observation: $\Sigma_\infty^D = \begin{bmatrix} V & C \\ C & V \end{bmatrix}$, by uniqueness of Σ_∞, we have $\frac{1}{2}(V + C) = \Sigma_\infty$.
  - $AMSE\left(\frac{\theta^A+\theta^B}{2}\right) = \frac{1}{2}\mathrm{Tr}\,(V + C) = AMSE(\theta)$.
  - $AMSE(\theta^A) = AMSE(\theta^B) > AMSE(\theta)$ due to $\mathrm{Tr}(V) \geq \mathrm{Tr}(C)$.

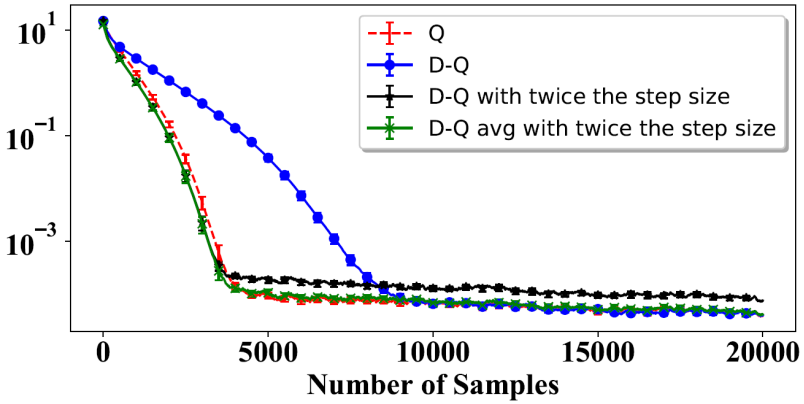
Baird's Example



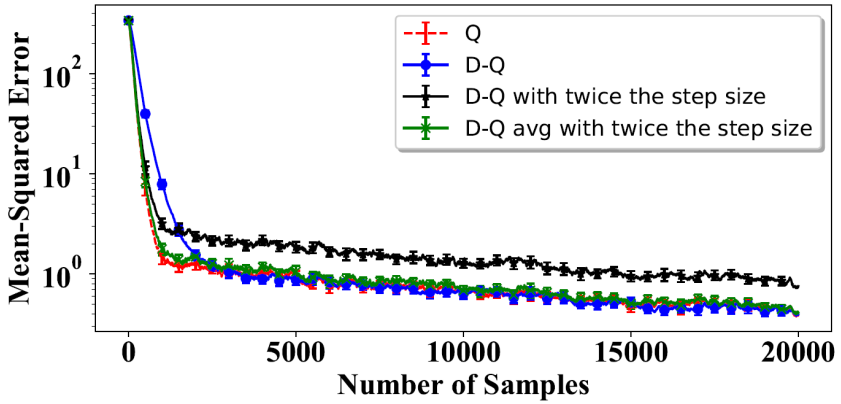
(a) Baird's Example



(b) Zero Reward



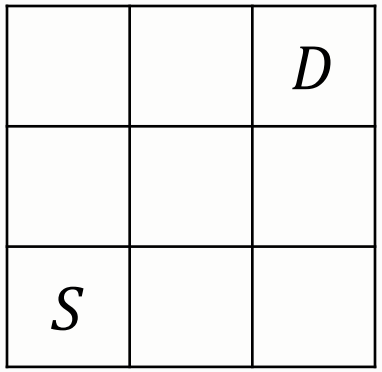
(c) Small Random Reward



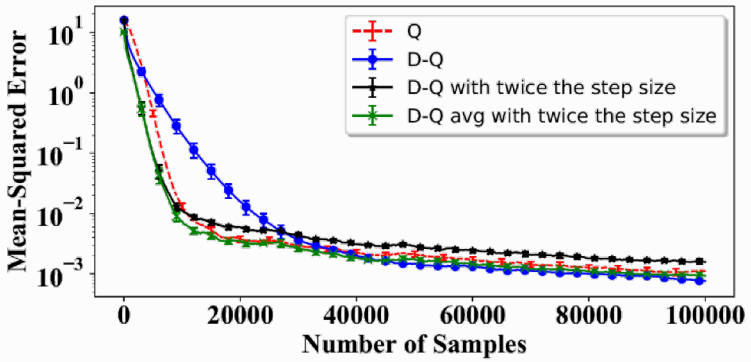
(d) Large Random Reward

Figure 1: Simulation results for Baird's example. The y-axis is in log scale.

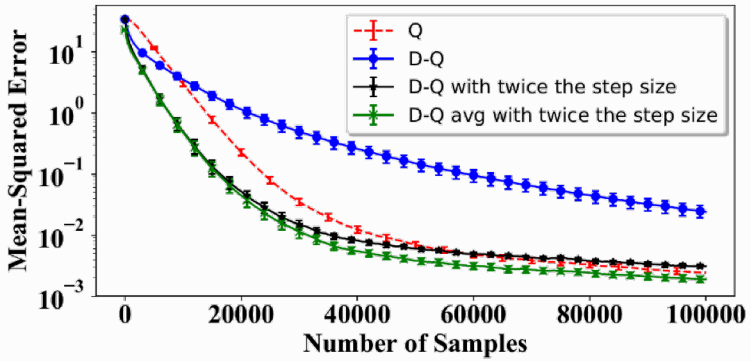
GridWorld



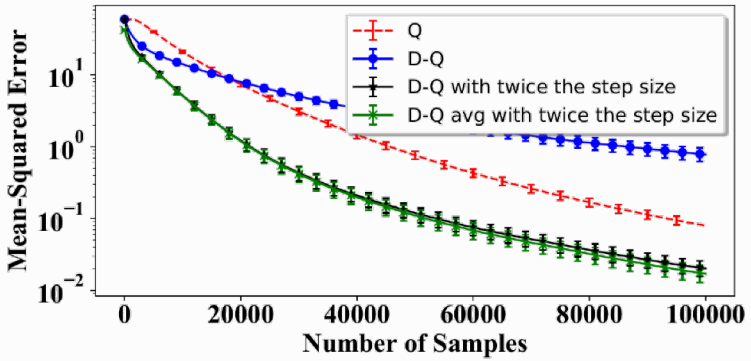
(a) An Example of 3×3 GridWorld



(b) 3×3 GridWorld



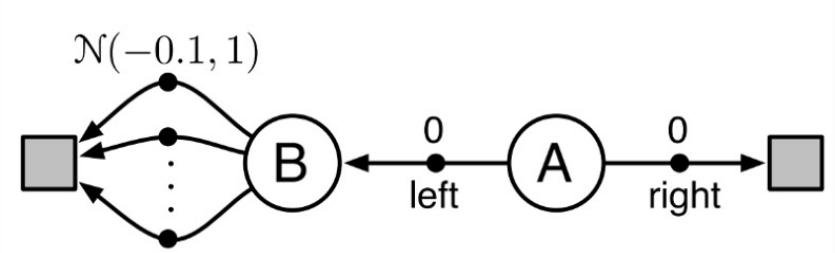
(c) 4×4 GridWorld

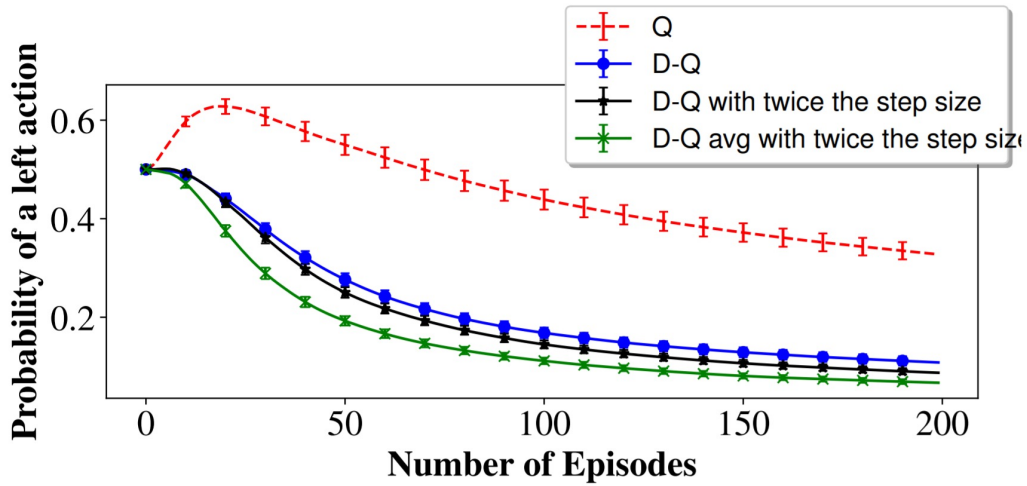


(d) 5×5 GridWorld

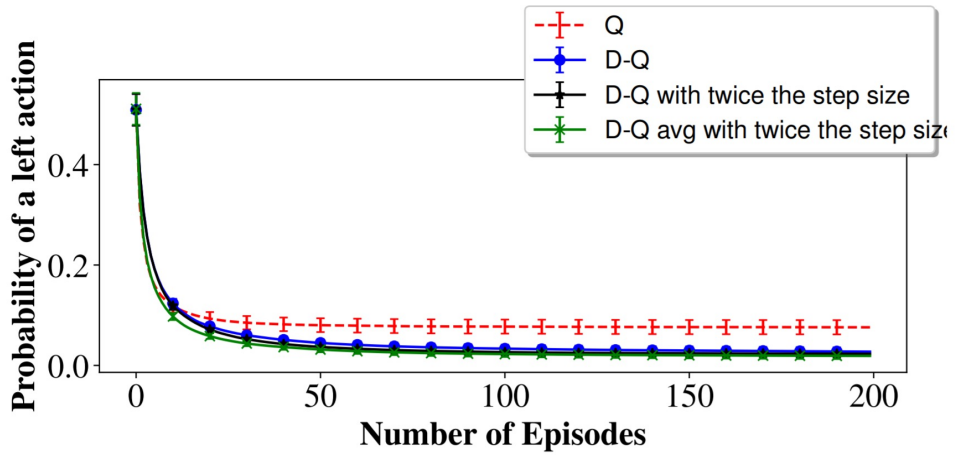
Figure 2: Simulation results for GridWorld with dimensions $3, 4, 5$. In all the three simulations, Double Q-learning with twice the step-size and averaged output outperforms Q-learning.

Sutton and Barto Example





(a) In a tabular setting with $M = 8$



(b) In a setting with neural network function approximations and $M = 10^9$

Observations

Is Double Q learning provably more efficient than Q-learning?

- Both from theoretical and numerical results:
  - Double Q-learning with the same stepsize converges slower than Q-learning;
  - Double Q-learning with twice stepsize can converge as fast as and even faster than Q-learning, but suffers from larger variance;
  - When using average estimator as the output, Double Q-learning with twice stepsize obtains both faster convergence rate and smaller mean-squared error.

From asymptotics to non-asymptotics

Asymptotic Convergence

- **TD-learning with LFA**
  [Tsitsiklis & Van Roy, 1997]
- **Double TD-learning with LFA**
  [Lee & He, 2019]
- **Synchronous Q-learning**
  [Borkar & Meyn, 2000]
- **Asynchronous Q-learning**
  [Jaakkola et al.,1994][Tsitsiklis, 1994] [Lee & He, 2020]
- **Q-learning with LFA**
  [Melo, Meyn, & Ribeiro, 2008] [Lee & He, 2020]
- **Greedy-GQ algorithm**
  [Maei et al., 2010]

Finite-time Convergence

- **TD-learning with LFA**
  [Srikant & Ying, 2019]
  [Dalal et al., 2018] [Bhandari et al., 2019]
  [Lakshminarayanan & Szepesvári, 2018]
- **Neural TD-learning**
  [Cai et al, 2019; Cayci et al., 2021]
- **Asynchronous Q-learning**
  [Szepesvári, 1998][Even-Dar & Mansour, 2003]
  [Qu & Wierman, 2020] [Li et al., 2020]
- **Q-learning with LFA**
  [Chen et al., 2019] [Wang & Giannakis,2020]
- **Double Q-learning**
  [Xiong et al, 2020]
- **Neural Q-learning**
  [Xu and Gu, 2020]

Tight Error Bound

- **TD-learning with LFA**
  [Hu & Syed, 2019]
  [Devraj & Meyn, 2017]
  [Chen et al., 2020]
- **Q-learning & Relative Q-learning**
  [Devraj & Meyn, 2020]
- **Double Q-learning**
  [Weng et al., 2020]

Summary

Existing rich control theory can help

- Build theoretical convergence of value-based RL algorithms
- Provide unified framework, tight characterization and error bounds
- Potentially design principled, data-efficient, robust, and extensible RL algorithms

Open Questions

- Nonlinear function approximations?
- Policy gradient methods?
- Global optimality?

Reference

- A Unified Switching System Perspective and Convergence Analysis of Q-Learning Algorithms, NeurIPS 2020.
  (with Donghwan Lee)

- The Mean-squared Error of Double Q-learning, NeurIPS 2020.
  (with Wentao Weng, Harsh Gupta, Ying Lei, and R. Srikant)

- Target-based Temporal Difference Learning, ICML 2019.
  (with Donghwan Lee)