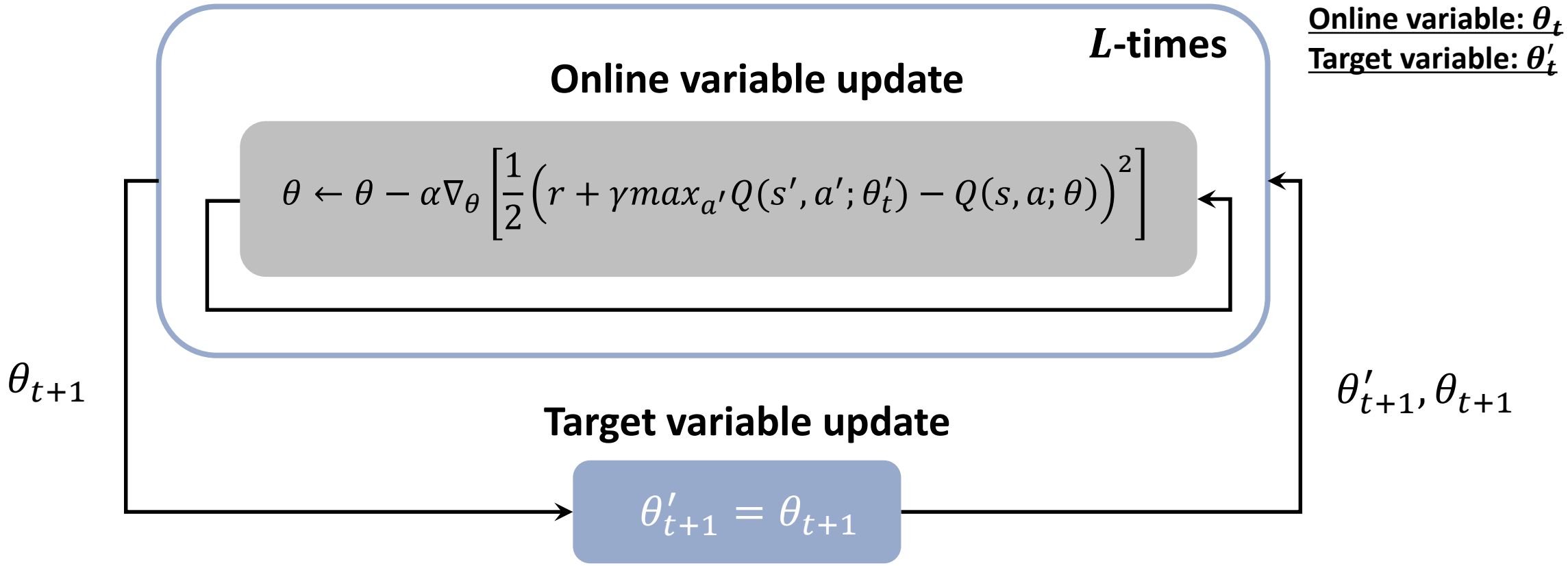
Target-Based Temporal-Difference Learning

Presenter: Niao He

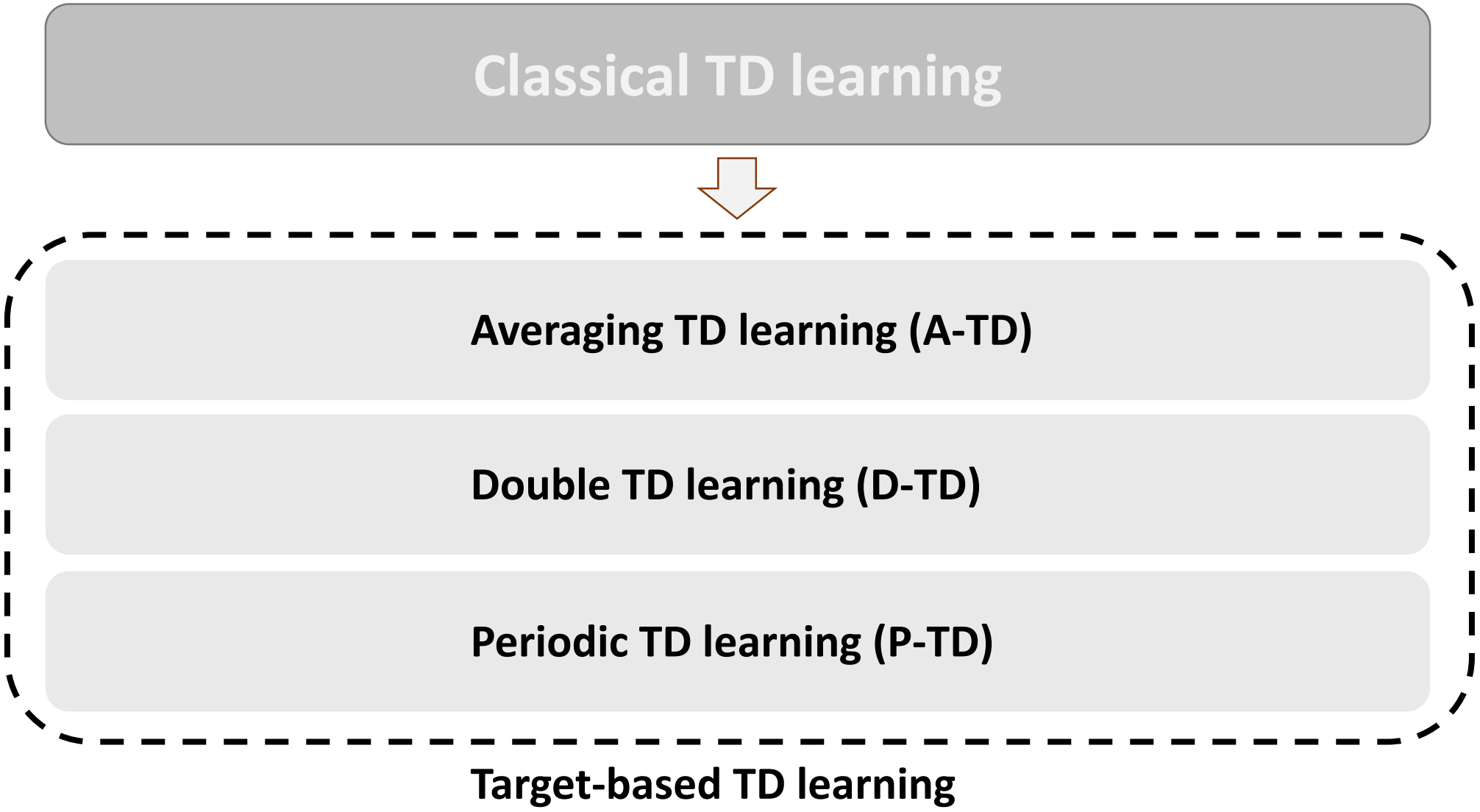
Donghwan Lee and Niao He
University of Illinois at Urbana-Champaign
ICML2019, Long Beach, CA

Deep Q-learning with target network

- The use of target network is pervasive in DQN-like algorithms.
- However, little is known from the theory side.



Temporal difference learning (TD learning)



Classical TD learning

Online variable update

$$\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta L(\theta; \theta'_t; e)\Big|_{\theta=\theta_t}$$

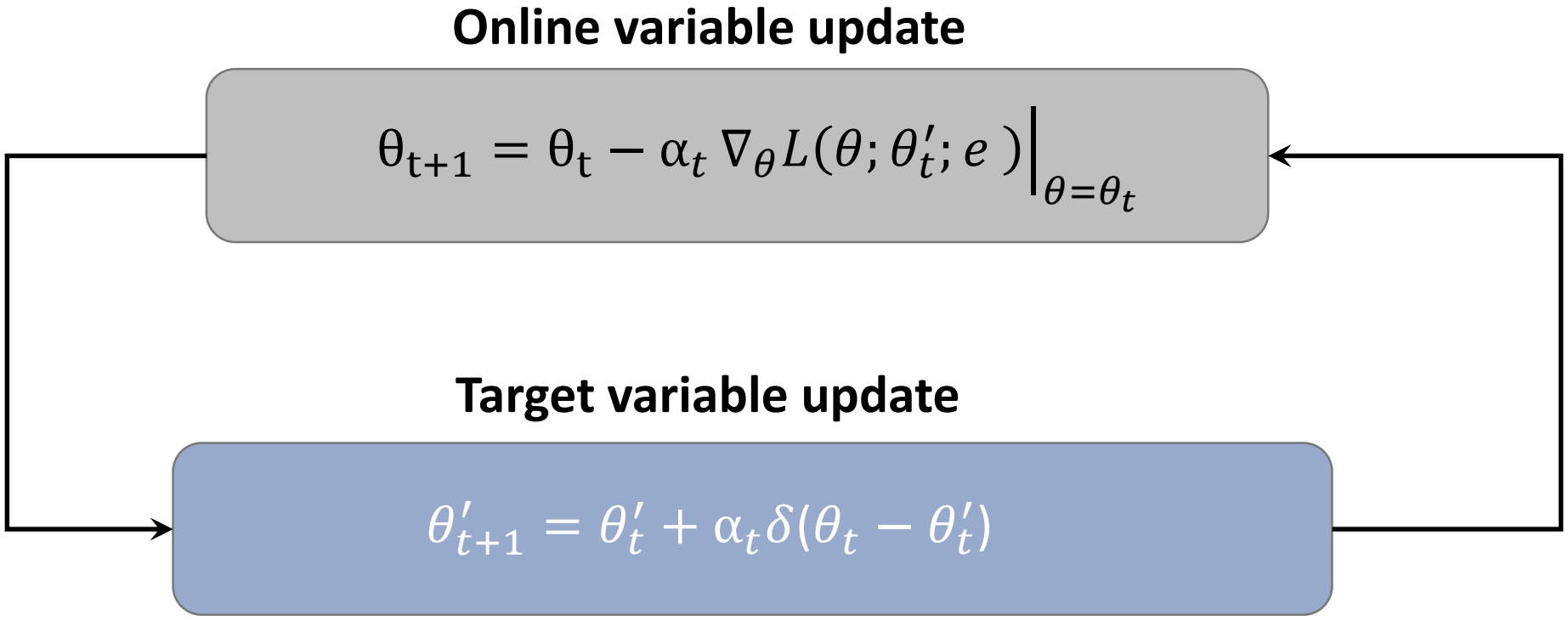
Target variable update

$$\theta'_{t+1} = \theta_{t+1}$$

Loss function of Bellman error

$$L(\theta; \theta'; e) := \frac{1}{2}\left(r - \gamma V(s'; \theta') - V(s; \theta)\right)^2, \qquad e := (s, a, r, s')$$

Averaging-TD learning (A-TD)



✓ Less aggressive target variable update by Polyak's averaging

Double TD learning (D-TD)

Online variable update

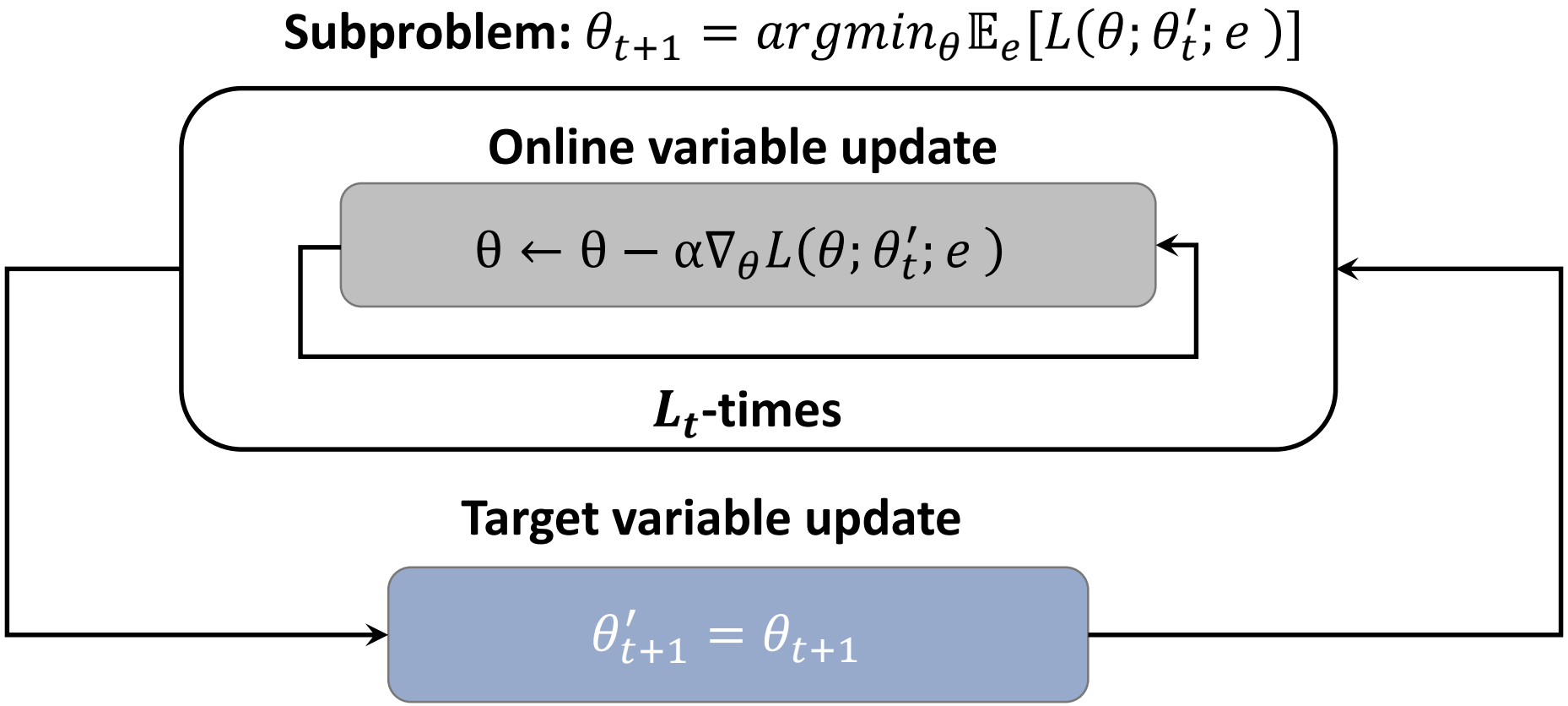
$$\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta \left(L(\theta; \theta'_t; e) + \frac{\delta}{2} \|\theta - \theta'_t\|_2^2 \right) \Big|_{\theta=\theta_t}$$

Target variable update

$$\theta'_{t+1} = \theta'_t - \alpha_t \nabla_{\theta'} \left(L(\theta'; \theta_t; e) + \frac{\delta}{2} \|\theta' - \theta_t\|_2^2 \right) \Big|_{\theta'=\theta'_t}$$

✓ Symmetrize the target and online updates

Periodic TD learning (P-TD)



✓ Take stochastic gradient steps L times in the inner loop

Convergence

Theorem: For A-TD and D-TD, $\theta_t \to \theta^*$ and $\theta'_t \to \theta^*$ as $t \to \infty$ with probability one, where θ^* is the solution of the projected Bellman equation

$$\Phi\theta = \Pi(R^\pi + \gamma P^\pi \Phi\theta)$$

and Π is the projection onto the range space of Φ

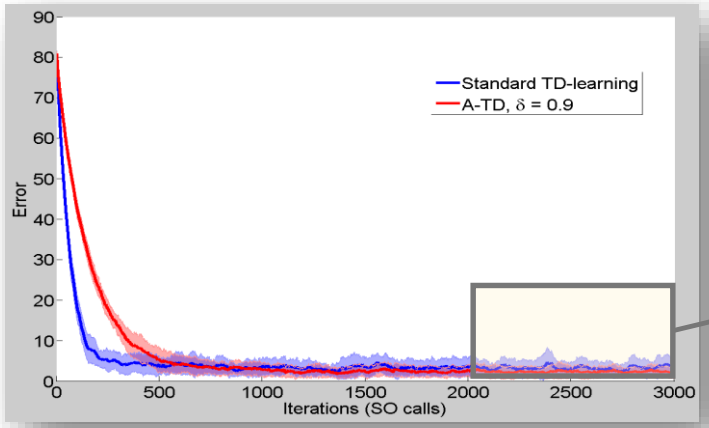
✓ The proof is based on the ODE and stochastic approximation

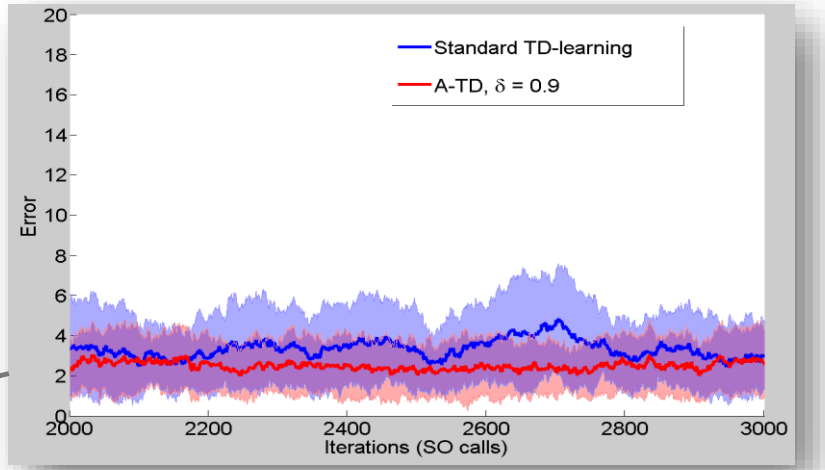
Theorem: For P-TD, an ϵ-optimal solution, $\mathbb{E}[\|\theta^* - \theta_t\|_D] \leq \epsilon$, is obtained by P-TD with at most $O\left(\left(\frac{1}{\epsilon^2}\right) ln\left(\frac{1}{\epsilon}\right)\right)$ samples.

✓The proof is based on standard results in stochastic gradient decent methods

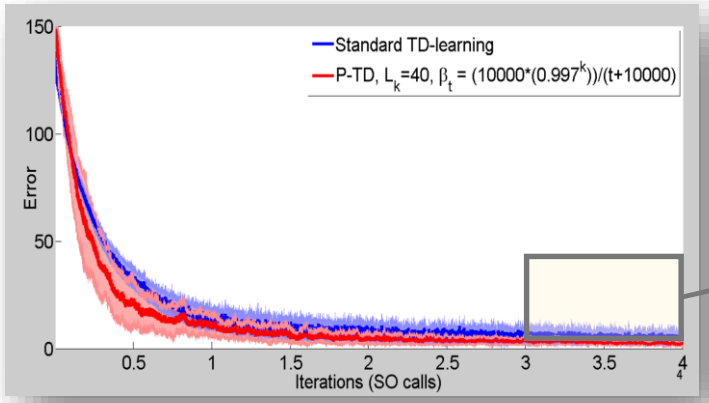
Simulations

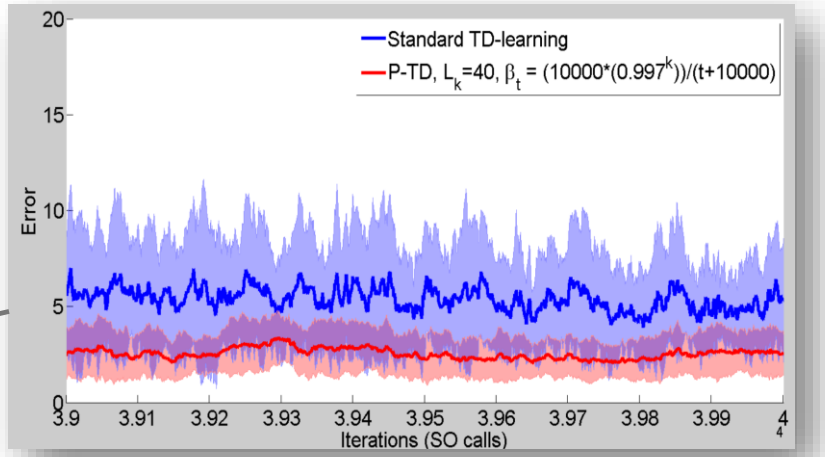
Standard TD
vs. A-TD





Standard TD
vs. P-TD





✓ After certain iterations, the target-based TD algorithms tend to show better convergence with lower variances.

Summary

✓Poster: Thu Jun 13th 06:30 -- 09:00 PM @ Pacific Ballroom #38