Two Time-scale Off-Policy TD Learning: Non-asymptotic Analysis over Markovian Samples

Tengyu Xu *; Shaofeng Zou †; Yingbin Liang *; *The Ohio State University, † University at Buffalo, The State University of New York

Introduction and Motivation

- **Markov Decision Process (MDP)** $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \pi, \mu_\pi)$:
  - $\mathcal{S}$: state space, $\mathcal{A}$: action space.
  - $\mathcal{P}(s'|s,a)$: transition kernel, $r(s,a,s')$: reward function.
  - $\gamma \in (0,1)$: discount factor.
  - $\pi(a|s)$: policy, i.e. conditional probability of choosing action a under state s.
  - μ_π: stationary distribution, i.e. $\sum_s p(s'|s)\mu_\pi(s) = \mu_\pi(s')$.
- **Off-policy value function evaluation**:
  - Value function: $v^\pi(s) = \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t, a_t, s_{t+1})|s_0 = s, \pi]$.
  - Goal: obtain the value function of target policy π given sample trajectory generated by behavior policy π_b.
  - Linear function approximation: using a linear function $\hat{v}(s,\theta) = \phi(s)\theta$ to approximate $v^\pi(s)$.
  - Challenge: vanilla TD learning could diverge in off-policy setting.

TDC Algorithm and Open Issues

- **TD with gradient correction (TDC) (R. Sutton (2009)):** minimizing mean-square projected Bellman error:

$$J(\theta) = \mathbb{E}_{\mu_{\pi_b}}[\hat{v}(s,\theta) - \Pi T^\pi \hat{v}(s,\theta)]^2.$$

  ▶ Global minimizer: $J(\theta^*) = 0$.
- **Two time-scale TDC update:**

$$\theta_{t+1} = \Pi_{R_\theta}(\theta + \alpha_t(A_t\theta_t + b_t + B_t w_t)),$$
$$w_{t+1} = \Pi_{R_w}(w_t + \beta_t(A_t\theta_t + b_t + C_t w_t)),$$

  - $A_t = \rho(s_t,a_t)\phi(s_t)(\gamma\phi(s_{t+1}) - \phi(s_t))$, $B_t = -\gamma\rho(s_t,a_t)\phi(s_{t+1})\phi(s_t)^\top$, $C_t = -\phi(s_t)\phi(s_t)^\top$ and $b_t = \rho(s_t,a_t)r(s_t,a_t,s_{t+1})\phi(s_t)$.
  - $\Pi_R(x) = \mathrm{argmin}_{x':\|x'\|_2 \le R}\|x - x'\|_2$ is the projection operator.
  - $R_\theta \ge \|A\|_2\|b\|_2$ and $R_w \ge 2\left\|C^{-1}\right\|_2\|A\|_2 R_\theta$.
  - $\rho(s,a) = \pi(s,a)/\pi_b(s,a)$ is the importance weighting factor.
- **Previous work:** G. Dalal et al.(2018): Two time-scale TDC under diminishing stepsize with i.i.d. samples satisfies:

$$\|\theta_t - \theta^*\|_2 = \mathcal{O}(t^{-2/3}) \text{ with high probability.}$$

- **Open issues:**
  - Convergence rate of TDC under diminishing stepsize with Markovian samples.
  - Convergence rate and convergence error of TDC under constant stepsize.
  - New update scheme for TDC that converges fast with small convergence error.

Technical Assumptions

- **Problem solvability:** $A = \mathbb{E}_{\mu_{\pi_b}}[\rho(s,a)\phi(s)(\gamma\phi(s') - \phi(s))^\top]$ and $C = -\mathbb{E}_{\mu_{\pi_b}}[\phi(s)\phi(s)^\top]$ are non-singular.
- **Bounded feature:** $\|\phi(s)\|_2 \le 1$ for all $s \in \mathcal{S}$ and $\rho_{\max} < \infty$.
- **Geometric ergodicity:** There exist constants $m > 0$ and $\rho \in (0,1)$ such that

$$\sup_{s\in\mathcal{S}} d_{TV}(\mathbb{P}(s_t \in \cdot|s_0 = s), \mu_{\pi_b}) \le m\rho^t, \forall t \ge 0,$$

where $d_{TV}(P,Q)$ denotes the total-variation distance between P and Q.

Contribution 1: Non-asymptotic Analysis under Diminishing Stepsize

Theorem 1. *Considering the diminishing stepsize $\alpha_t = \mathcal{O}(t^{-\sigma})$ and $\beta_t = \mathcal{O}(t^{-\nu})$. If $0 < \nu < \sigma < 1$, the output of two time-scale TDC satisfies*

$$\mathbb{E}\|\theta_t - \theta^*\|_2^2 = \mathcal{O}\Big(\frac{\log t}{t^\nu} + h(\sigma,\nu)\Big)$$

where

$$h(\sigma,\nu) = \begin{cases} \frac{\log t}{t^\nu}, & \sigma > 1.5\nu, \\ \frac{1}{t^{2(\sigma-\nu)}}, & \sigma \le 1.5\nu. \end{cases}$$

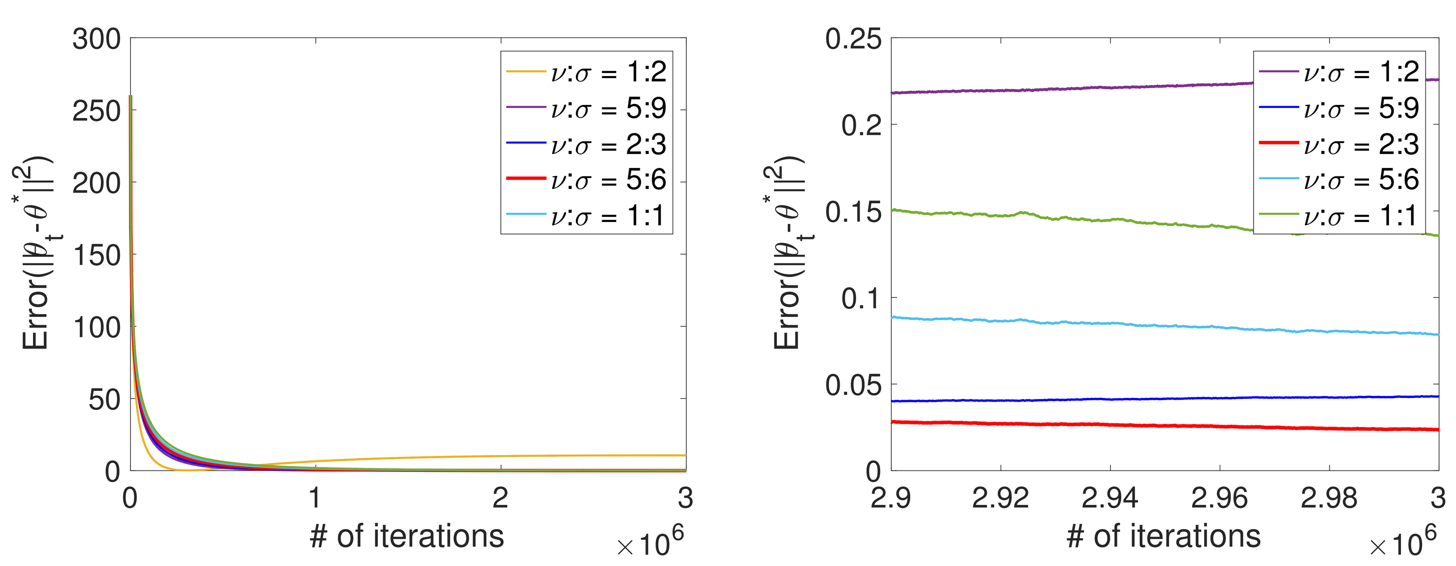
If $0 < \nu < \sigma = 1$, then we have

$$\mathbb{E}\|\theta_t - \theta^*\|_2^2 = \mathcal{O}\Big(\frac{\log^2 t}{t^\nu} + h(1,\nu)\Big).$$

– The optimal convergence is obtained when $\sigma = 1.5\nu$, with $\sigma = 1$ yields the best error decay rate $\mathcal{O}(\log^2 t/t^{2/3})$.

- **Proof Sketch**
  - Formulate update of training error $\|\theta_t - \theta^*\|_2$ and tracking error $\|z_t\|_2 = \left\|w_t + C^{-1}(b + A\theta_t)\right\|_2$.
  - Derive preliminary bound of tracking error: $\mathbb{E}\|z_t\|_2^2 = \mathcal{O}(t^{-(\sigma-\nu)})$.
  - Recursively refine bound of tracking error: $\mathbb{E}\|z_t\|_2^2 = \mathcal{O}(\log t/t^\nu + h(\sigma,\nu))$.
  - Derive bound of $\mathbb{E}\|\theta_t - \theta^*\|_2^2$ based on the bound of $\mathbb{E}\|z_t\|_2^2$.

Optimal Diminishing Stepsize



Comparison among diminishing stepsize with $\sigma = 0.6$. (left: full; right: tail)

- Diminishing stepsize satisfying $\sigma = 1.5\nu$ yields the best error decay rate.

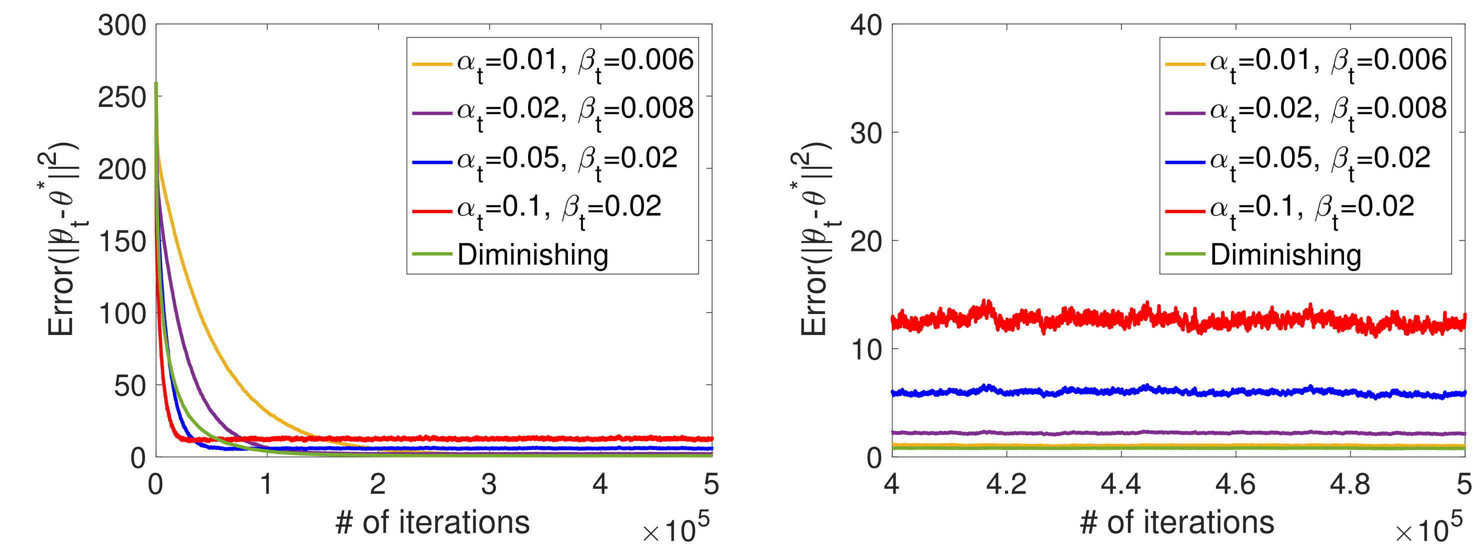
Contribution 2: Non-asymptotic Analysis under Constant Stepsize

Theorem 2. *Considering the constant stepsize $\alpha_t = \alpha$, $\beta_t = \beta$. The output of two time-scale TDC satisfies:*

$$\mathbb{E}\|\theta_t - \theta^*\|_2^2 = \mathcal{O}((1 - |\lambda_\theta|\alpha)^t) + \mathcal{O}(\max\{\alpha, \alpha\ln\frac{1}{\alpha}\}) + \mathcal{O}(\max\{\beta, \beta\ln\frac{1}{\beta}, \frac{\alpha}{\beta}\})^{0.5}$$

- Converges fast to a neighborhood of θ^* at a linear rate when α is large.
- Large α, β and α/β cause large training error.

Optimal Constant Stepsize



Comparison between TDC updates under constant stepsizes and diminishing stepsize. (left: full; right: tail)

- TDC with large constant stepsize converges fast but has large training error

Contribution 3: Blockwise Diminishing Stepsize

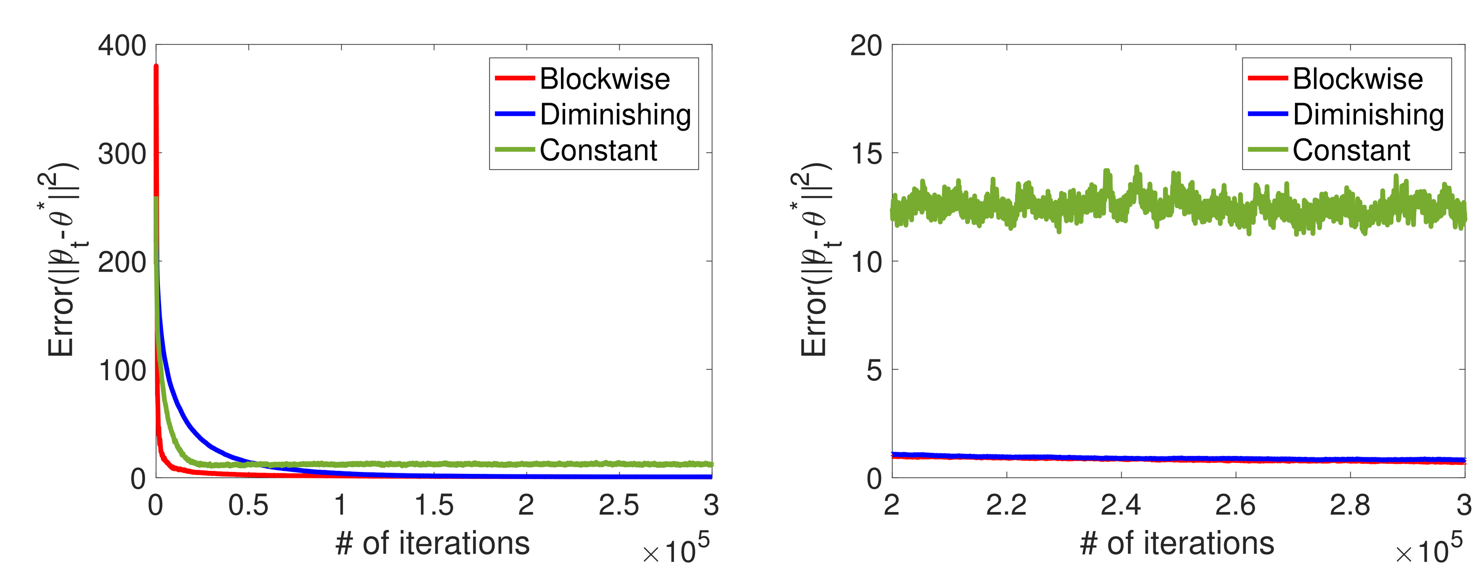
Key idea: α_s and β_s are kept constant within each block with length T_s and diminished blockwisely.

Theorem 3. *Suppose* $\max\{\log(1/\alpha_s)\alpha_s, \alpha_s\} \le \|\theta_0 - \theta^*\|_2/2^s$, $\alpha_s/\beta_s \ge 1/2\max\{0, \lambda_{\min}(C^{-1}(A^\top + A))\}$ *and* $T_s = \lceil \log_{1/(1-|\lambda_x|\alpha_s)} 4 \rceil$, *where λ_x is a constant. Then, after $S = \lceil \log_2(\|\theta_0 - \theta^*\|_2/\epsilon) \rceil$ blocks, we have*

$$\mathbb{E}\|\theta_S - \theta^*\|_2^2 \le \epsilon.$$

The total sample complexity is $\mathcal{O}(\frac{1}{\epsilon}\log^2(\frac{1}{\epsilon}))$.

Comparison between Different Stepsizes



Comparison between TDC updates under blockwise diminishing stepsizes, diminishing stepsize and constant stepsize. (left: full; right: tail)

- TDC under blockwise diminishing stepsize converges faster than that under diminishing stepsize and almost as fast as that under constant stepsize.
- TDC under blockwise diminishing stepsize has comparable training error as that under diminishing stepsize