Sample Complexity Bounds for Two Timescale Value-based Reinforcement Learning Algorithms

Tengyu Xu
The Ohio State University
xu.3260@osu.edu

Yingbin Liang
The Ohio State University
liang.889@osu.edu

Abstract

Two timescale stochastic approximation (SA) has been widely used in value-based reinforcement learning algorithms. In the policy evaluation setting, it can model the linear and nonlinear temporal difference learning with gradient correction (TDC) algorithms as linear SA and nonlinear SA, respectively. In the policy optimization setting, two timescale nonlinear SA can also model the greedy gradient-Q (Greedy-GQ) algorithm. In previous studies, the non-asymptotic analysis of linear TDC and Greedy-GQ has been studied in the Markovian setting, with diminishing or accuracy-dependent stepsize. For the nonlinear TDC algorithm, only the asymptotic convergence has been established. In this paper, we study the non-asymptotic convergence rate of two timescale linear and nonlinear TDC and Greedy-GQ under Markovian sampling and with accuracy-independent constant stepsize. For linear TDC, we provide a novel non-asymptotic analysis and show that it attains an $\epsilon$-accurate solution with the optimal sample complexity of $O(\epsilon^{-1}\log(1/\epsilon))$ under a constant stepsize. For nonlinear TDC and Greedy-GQ, we show that both algorithms attain $\epsilon$-accurate stationary solution with sample complexity $O(\epsilon^{-2})$. It is the first non-asymptotic convergence result established for nonlinear TDC under Markovian sampling and our result for Greedy-GQ outperforms the previous result orderwisely by a factor of $O(\epsilon^{-1}\log(1/\epsilon))$.

1 Introduction

Two timescale stochastic approximation (SA) algorithms have wide applications in reinforcement learning (RL) [Sutton and Barto, 2018]. Typically, two timescale SA algorithms involve iterations of two types of variables updated at different speeds, i.e., the stepsizes for two iterates are chosen differently so that one iterate runs much faster than the other [Borkar, 1997, Borkar, 2009]. Such algorithms are widely used to solve both policy evaluation and policy optimization problems in RL, in which the goal of policy evaluation is to estimate the expected total reward (i.e., value function) of a target policy, and the goal of policy optimization is to search for a policy with the optimal expected total reward.

In the policy evaluation problem, temporal difference (TD) learning [Sutton, 1988] is one of the most widely used algorithms when a linear function class is utilized to approximate the value function. However, in the off-policy setting, in which the target policy to be evaluated is different from the behavior policy that generates samples, TD learning may diverge to infinity. To overcome such an issue, [Sutton et al., 2009] proposed the two timescale linear TD with gradient correction (TDC) algorithm, which has convergence guarantee in the off-policy setting. The two timescale linear TDC is a special case of two timescale linear SA, whose asymptotic convergence has been established in [Sutton et al., 2009, Borkar, 2009] and [Yu, 2017, Tadic, 2004, Yaji and Bhatnagar, 2016] for the i.i.d. and Markovian settings, respectively. The non-asymptotic convergence rate of two timescale linear TDC/SA has also been studied. In the i.i.d. setting, under diminishing stepsize, [Dalal et al., 2018b] established the sample complexity of $O(\epsilon^{-1.5})$, and an improved complexity of $O(\epsilon^{-1})$ was later established in [Dalal et al., 2019]. In the Markovian setting, [Xu et al., 2019] established the complexity of $O(\epsilon^{-1.5}\log^2(1/\epsilon))$ under a diminishing stepsize, and [Gupta et al., 2019] established the complexity of $O(\epsilon^{-1}\log^2(1/\epsilon))$ under a $\epsilon$-dependent stepsize, where
\(\zeta\) can be an arbitrarily small positive constant. Recently, [Kaledin et al., 2020] provides a tighter complexity bound of \(O(\epsilon^{-1})\) for two timescale linear SA under a diminishing stepsize. Although having progressed significantly, existing convergence guarantee were established either under a diminishing stepsize or a drastically small \(\epsilon\)-level stepsize, which yield very slow convergence and are rarely used in practice.

- **Thus, the first goal of this paper is to investigate the two timescale linear TDC under a constant stepsize (not \(\epsilon\)-dependent), which is commonly adopted in practice, and to provide the finite-sample convergence guarantee for such a case. This necessarily requires a new approach differently from the existing ones.**

When a nonlinear function is utilized to approximate the value function, TD learning still suffers from the divergence issue [Tsitsiklis and Van Roy, 1997]. To address that, [Bhatnagar et al., 2009] proposed the two timescale nonlinear TDC, which can be modeled as a two timescale nonlinear SA. The asymptotic convergence of two timescale nonlinear SA has been well established in [Borkar, 1997] and [Tadic, 2004]. However, the non-asymptotic convergence of two timescale nonlinear SA has only been established in the i.i.d. setting under some restrict assumptions such as global (local) stability and local linearization [Borkar and Pattathil, 2018, Mokkadem and Pelletier, 2006]. So far, the non-asymptotic convergence performance of two timescale nonlinear TDC has not been studied under the general Markovian sampling.

- **The second goal of this paper is to provide the first non-asymptotic convergence analysis for two timescale nonlinear TDC with a constant stepsize, under Markovian sampling, and without restricted assumptions.**

Moreover, in the policy optimization problem, Q-learning [Watkins and Dayan, 1992] has been widely used and has achieved significant success in practice. However, in the function approximation setting, Q-learning does not have convergence guarantee [Baird, 1995] unless under some restricted regularity assumptions [Melo et al., 2008, Zou et al., 2019, Cai et al., 2019]. In corresponding to this, [Maei and Sutton, 2010] proposed the Greedy-GQ algorithm in the linear function approximation setting, in which the algorithm is guaranteed to converge to a locally optimal policy without restricted assumptions. Similarly to nonlinear TDC algorithms, Greedy-GQ also adopts a two timescale update scheme, and is a special case of two timescale nonlinear SA. Under single-sample update and Markovian sampling, [Wang and Zou, 2020] provided the non-asymptotic convergence rate of Greedy-GQ with diminishing stepsize, which achieves the complexity of \(O(\epsilon^{-3}\log(\epsilon^{-1}))\). However, such a rate does not attain the typical complexity order of nonconvex optimization, and can be potentially improved with a larger stepsize.

- **The last focus of this paper is to provide an improved non-asymptotic convergence rate for two timescale Greedy-GQ under a constant stepsize.**

### 1.1 Our Contributions

For two timescale linear TDC, we show that it achieves the sample complexity of \(O(\epsilon^{-1}\log(\epsilon^{-1}))\), which has the optimal dependence on \(\epsilon\) due to the lower bound given in [Dalal et al., 2019]. Such a rate has been established in [Kaledin et al., 2020], but only under a diminishing stepsize, which is rarely used in practice due to the slow empirical performance. In contrast, our guarantee is established under a constant (not \(\epsilon\)-dependent) stepsize, which is commonly used in practice. Our analysis approach leverages the mini-batch sampling for each iteration to control the convergence error, which is significantly different from that in [Kaledin et al., 2020], and can be of independent interest.

For two timescale nonlinear TDC, we establish the first non-asymptotic convergence rate under Markovian sampling. We show that the mini-batch two timescale nonlinear TDC algorithm achieves the sample complexity of \(O(\epsilon^{-2})\).

For two timescale Greedy-GQ, we show that mini-batch two timescale Greedy-GQ with a constant stepsize and under Markovian sampling achieves the sample complexity of \(O(\epsilon^{-2})\). Our result orderwisely outperforms the previous result of Greedy-GQ with diminishing stepsize in [Wang and Zou, 2020] by a factor of \(O(\epsilon^{-1}\log(1/\epsilon))\).

### 1.2 Related Work

Due to the vast amount of studies on SA and value-based RL algorithms, we include here only the studies that are highly related to our work.

**Policy evaluation with linear function approximation.** In the on-policy setting, TD learning [Sutton, 1988] has been proposed to solve the policy evaluation problem in the linear function approximation setting. The non-asymptotic convergence rate of TD learning has been established in [Dalal et al., 2018a] for the i.i.d. setting and in [Bhandari et al., 2018, R. Srikant, 2019]. [Hu and Syed, 2019] for the Markovian setting. [Cai et al., 2019] explored the linearizable structure...
of neural networks in the overparameterized regime, and studied the non-asymptotic convergence rate of TD learning with neural network approximation. [Zou et al., 2019] studied the convergence rate of SARSA with linear function approximation in the Markovian setting, which can be viewed as a policy evaluation with dynamic changing transition kernel.

In the off-policy setting, GTD, GTD2 and TDC have been proposed to solve the divergence issue of TD learning [Sutton et al., 2008] Sutton et al., 2009 Maae, 2011]. The convergence rate of one timescale GTD and GTD2 algorithms has been established in [Liu et al., 2015] by converting the objective into a convex-concave saddle problem in the i.i.d. setting, and was further generalized to the Markovian setting in [Wang et al., 2017]. For two timescale linear TDC, in the i.i.d. setting, the non-asymptotic analysis was provided in [Dalal et al., 2018b] Dalal et al., 2019]. In the Markovian setting, the non-asymptotic convergence rate was first established in [Xu et al., 2019] under diminishing stepsize and in [Gupta et al., 2019] under constant stepsizes. The result in [Xu et al., 2019] was later improved by [Kaledin et al., 2020] to achieve the optimal convergence rate.

Policy evaluation with nonlinear function approximation. Two timescale nonlinear TDC is proposed by [Bhatnagar et al., 2009], in which a smooth nonlinear function is utilized to approximate the value function. Nonlinear TDC with i.i.d. samples is a special case of two time-scale nonlinear SA with martingale noise, whose asymptotic convergence has been established in [Bhatnagar et al., 2009] [Maei, 2011] by using asymptotic convergence results in nonlinear SA [Borkar, 1997] Borkar, 2009 Tadic, 2004]. Under the global/local asymptotic stability assumptions or local linearization assumption, the non-asymptotic convergence of two timescale nonlinear SA with martingale noise has been studied in [Borkar and Pattathil, 2018]. Under certain stability assumptions, the asymptotic convergence of two timescale nonlinear SA with Markov noise was established in [Karmakar and Bhatnagar, 2018]. A concurrent study [Qiu et al., 2020] also investigated nonlinear TDC and obtained the same sample complexity of $O(\epsilon^{-2})$ as our result. However, [Qiu et al., 2019] only considered the i.i.d. setting, whereas we considered the more general Markovian setting.

Policy optimization with linear function approximation. Q-learning [Watkins and Dayan, 1992] is one of the most widely used value-based policy optimization algorithms. The asymptotic and non-asymptotic convergence have been established for Q-learning with linear function approximation in [Melo et al., 2008] and [Zou et al., 2019], respectively, under certain regularity assumption. Under a similar regularity assumption, [Cai et al., 2019] established the convergence rate of Q-Learning in the neural network approximation setting. However, without regularity assumptions, Q-Learning does not have convergence guarantee in the function approximation setting. [Maei et al., 2010] proposed two timescale Greedy-GQ to solve the divergence issue of Q-Learning with linear function approximation, and the asymptotic convergence of Greedy-GQ was also established therein. Recently, [Wang and Zou, 2020] studied the non-asymptotic convergence rate of Greedy-GQ under diminishing stepsize in the Markovian setting. In this paper, we provide an orderwisely better convergence rate than that in [Wang and Zou, 2020].

2 Markov Decision Process

Consider a Markov decision process (MDP) denoted $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Here, $\mathcal{S} \subset \mathbb{R}^d$ is a state space, $\mathcal{A}$ is an action set, $P = P(s'|s,a)$ is the transition kernel, $r(s,a,s')$ is the reward function bounded by $r_{\text{max}}$, and $\gamma \in (0, 1)$ is the discount factor. A stationary policy $\pi$ maps a state $s \in \mathcal{S}$ to a probability distribution $\pi(\cdot|s)$ over the action space $\mathcal{A}$. At time-step $t$, suppose the process is in some state $s_t \in \mathcal{S}$. Then an action $a_t \in \mathcal{A}$ is taken based on the distribution $\pi(\cdot | s_t)$, the system transitions to a next state $s_{t+1} \in \mathcal{S}$ governed by the transition kernel $P(\cdot|s_t,a_t)$, and a reward $r_t = r(s_t,a_t,s_{t+1})$ is received. We assume the associated Markov chain $p(s'|s,a) = \sum_{a \in \mathcal{A}} P(s'|s,a)\pi(a|s)$ is ergodic, and let $\mu_\pi$ be the induced stationary distribution of this MDP, i.e., $\sum_{s} p(s'|s)\mu_\pi(s) = \mu_\pi(s')$. The state value function for policy $\pi$ is defined as: $V^\pi(s) = \mathbb{E}_{t=0}^\infty \gamma^t r(s_t,a_t,s_{t+1})|s_0 = s, \pi]$, and the state-action value function is defined as: $Q^\pi(s,a) = \mathbb{E}_{t=0}^\infty \gamma^t r(s_t,a_t,s_{t+1})|s_0 = s, a_0 = a, \pi$. It is known that $V^\pi(s)$ is the unique fixed point of the Bellman operator $T^\pi$, i.e., $V^\pi(s) = T^\pi V^\pi(s) = r^\pi(s) + \gamma \mathbb{E} s', s' \sim T^\pi(s') V^\pi(s')$, where $r^\pi(s) = \mathbb{E}_{a,s} r(s,a,s')$ is the expected reward of the Markov chain induced by the policy $\pi$. We take the following standard assumption for the MDP in this paper, which has also been adopted in previous works [Bhandari et al., 2018] Zou et al., 2019 R. Srikant, 2019 Xu et al., 2019 Xu et al., 2020b] .

Assumption 1 (Geometric ergodicity). There exist constants $\kappa > 0$ and $\rho \in (0, 1)$ such that

$$\sup_{s \in \mathcal{S}} d_{TV}(\mathbb{P}(s_t|s_0 = s), \mu_\pi(s_t)) \leq \kappa \rho^t, \forall t \geq 0,$$

where $\mathbb{P}(s_t|s_0 = s)$ is the distribution of $s_t$ conditioned on $s_0 = s$ and $d_{TV}(P, Q)$ denotes the total-variation distance between the probability measures $P$ and $Q$.

Assumption 1 holds for any time-homogeneous Markov chain with finite state space and any
uniformly ergodic Markov chain with general state space [Bhandari et al., 2018; Zou et al., 2019; Xu et al., 2019].

3 Two Timescale TDC with Linear Function Approximation

In this section we first introduce the two timescale linear TDC algorithm to solve the policy evaluation problem, and then present our convergence rate result.

3.1 Algorithm

When $S$ is large or infinite, a linear function $\hat{v}(s, \theta) = \phi(s)\theta^T$ is often used to approximate the value function $V^\pi(s)$, where $\phi(s) \in \mathbb{R}^d$ is a fixed feature vector for state $s$ and $\theta \in \mathbb{R}^d$ is a parameter vector. We can also write the linear approximation in the vector form as $\hat{v}(\theta) = \Phi \theta$, where $\Phi$ is the $|S| \times d$ feature matrix. Without loss of generality, we assume that the feature vector $\|\phi(s)\|_2 \leq 1$ for all $s \in S$ and the columns of the feature matrix $\Phi$ are linearly independent. Here we consider policy evaluation problem in the off-policy setting. Namely, a sample path $\{(s_1, a_1, s_{t+1})\}_{t \geq 0}$ is generated by the Markov chain according to a behavior policy $\pi_b$, but our goal is to obtain the value function of a target policy $\pi_t$, which is different from $\pi_b$.

To find a parameter $\theta^* \in \mathbb{R}^d$ with $\mathbb{E}_{\mu_b} \hat{v}(s, \theta^*) = \mathbb{E}_{\mu_b} T^\pi \hat{v}(s, \theta^*)$. The linear TDC algorithm [Sutton et al., 2009] updates the parameter by minimizing the mean-square projected Bellman error (MSPBE) objective, defined as

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

where $\Pi$ is the orthogonal projection operation onto the function space $\hat{V} = \{ \hat{v}(\theta) \mid \theta \in \mathbb{R}^d \text{ and } \hat{v}(\cdot, \theta) = \phi(\cdot)\theta^T \}$. The columns of the feature matrix $\Phi$ are linearly independent, [Sutton et al., 2009] shows that $J(\theta)$ is strongly convex and has $\theta^* = -A^{-1}b$ as its global minimum, i.e., $J(\theta^*) = 0$, where $A = \mathbb{E}_{\mu_b} [\gamma \mathbb{E}_{\pi} [\phi(s')|s] - \phi(s)|s]$ and $b = \mathbb{E}_{\mu_b} [\mathbb{E}_{\pi} [r(s, a, s')|s]|s]$. A convenient way to find $\theta^*$ is to minimize the MSPBE objective function $J(\theta)$ using the gradient descent method: $\theta_{t+1} = \theta_t - \frac{1}{2} \nabla J(\theta_t)$, where $\theta_t \geq 0$ is the stepsize and the gradient $\nabla J(\theta)$ was derived by [Bhatnagar et al., 2009] as follows:

$$-\frac{1}{2} \nabla J(\theta) = \mathbb{E}_{\mu_b} [\mathbb{E}_{\pi}[\delta(\theta)|s]\phi(s)] - \gamma \mathbb{E}_{\mu_b} [\mathbb{E}_{\pi}[\phi(s')|s]\phi(s)^T] w(\theta),$$

where $\delta(\theta) = r(s, a, s') + \gamma \hat{v}(s', \theta) - \hat{v}(s, \theta)$ is the temporal difference error, $w(\theta) := \mathbb{E}_{\mu_b} [\phi(s)\phi(s)^T]^{-1} \mathbb{E}_{\mu_b} [\mathbb{E}_{\pi}[\delta(\theta)|s]\phi(s)]$. In practice, stochastic gradient descent (SGD) method is usually adopted to perform the update in eq. (1) approximately. However, directly sampling is not applicable to $w(\theta)$. To solve such an issue, an auxiliary parameter $w_t$ can be introduced to estimate the vector $w(\theta_t)$, i.e., $w_t \approx w(\theta_t)$, by solving a linear SA with the following corresponding ODE:

$$\dot{w} = -\mathbb{E}_{\mu_b} [\phi(s)\phi(s)^T] w + \mathbb{E}_{\mu_b} [\mathbb{E}_{\pi}[\delta(\theta)|s]\phi(s)].$$

Given $w_t$, the parameter $\theta_t$ can then be updated with a stochastic approximation of $\nabla J(\theta_t)$ obtained via directly sampling:

$$\theta_{t+1} = \theta_t + \alpha \frac{1}{|B_t|} \sum_{j \in B_t} g(\theta_t, w_{t_j}, x_{t_j}),$$

where $B_t$ is the mini-batch sampled from the MDP, $g(\theta_t, w_{t_j}, x_{t_j}) = \rho(s_j, a_j)(\delta_j(\theta_t)\phi(s_j) - \gamma \phi(s_{j+1})\phi(s_j)^T w_t)$, $\rho(s, a) = \pi_b(a|s)/\pi_b(a|s)$ is the importance weighting factor with $\max \rho$ being its maximum value, and $x_j$ denotes the sample $(s_j, a_j, s_{j+1})$.

Algorithm II is an online algorithm based on a single sample path. Algorithm II adopts a two timescale update scheme, in which parameters $\theta_t$ and $w_t$ are updated simultaneously but with different stepizes. Specifically, the main parameter $\theta_t$ iterates at a slow timescale with a smaller stepsize, and the auxiliary parameter $w_t$ iterates at a fast timescale with a larger stepize. By doing so, $w_t$ can be close to $w(\theta_t)$ asymptotically, so that $\theta_t$ is updated approximately in the direction of $-\nabla J(\theta)$. Algorithm II utilizes an accuracy-independent constant stepsize, i.e., $\alpha, \beta = O(1)$ for both the updates of $\theta_t$ and $w_t$, and a mini-batch of samples $\{(s_j, a_j, s_{j+1})\}_{t \leq i < t+M-1}$ are taken sequentially from the trajectory at each iteration to perform the update. As we will show later, linear TDC in this setting is guaranteed to converge to the global optimal with an arbitrary accuracy level.

Algorithm 1 Two Timescale Linear TDC

1: **Input:** batch size $M$, learning rate $\alpha$ and $\beta$

2: **Sampling:** A trajectory $\{(s_j, a_j)\}_{t \geq 0}$ is sampled by following the behaviour policy $\pi_b$

3: **Initialization:** $\theta_0$ and $w_0$

4: for $t = 0, \cdots, T - 1$ do

5: $t_t = tM$

6: $w_{t_t + 1} = w_t + \beta \frac{1}{M} \sum_{j = t_t}^{t_t + M - 1} (-\phi(s_j)\phi(s_j)^T w_t + \rho(s_j, a_j)\delta_j(\theta_t)\phi(s_j))$

7: $\theta_{t_t + 1} = \theta_t + \alpha \frac{1}{M} \sum_{j = t_t}^{t_t + M - 1} \rho(s_j, a_j)(\delta_j(\theta_t)\phi(s_j) - \gamma \phi(s_{j+1})\phi(s_j)^T w_t)$

8: end for

9: **Output:** $\theta_T$
3.2 Convergence Analysis

We define matrix $C = -E_{\mu_\pi} [\phi(s)\phi(s)']$. Let $\lambda_1 = \max(A^T C^{-1} A)$, $\lambda_2 = \max(C)$, and $R_0 = ||\theta||_2$. The following theorem provides the convergence rate and sample complexity of Algorithm 1.

**Theorem 1.** Suppose Assumption 1 hold. Consider Algorithm 1 of two timescale linear TDC update. Let the stepsize $\alpha \leq \min\left\{ \frac{4}{\lambda_1^2}, \frac{\sqrt{2\lambda_2 \lambda_3}}{4\sqrt{\beta} \max\left\{ \frac{\lambda_2}{\sqrt{2\lambda_3}}, \frac{\lambda_2}{\sqrt{2\lambda_2}} \right\}}, \frac{\sqrt{\lambda_2 \lambda_3}}{4\sqrt{\beta} \max\left\{ \frac{\lambda_2}{\sqrt{2\lambda_3}}, \frac{\lambda_2}{\sqrt{2\lambda_2}} \right\}} \right\}$, $\beta \leq \min\left\{ \frac{4}{\lambda_1^2}, \frac{\lambda_2}{4\sqrt{\beta} \max\left\{ \frac{\lambda_2}{\sqrt{2\lambda_3}}, \frac{\lambda_2}{\sqrt{2\lambda_2}} \right\}} \right\}$ and the batch size $M \geq 128 \left( \rho_{\text{max}}^2 + \frac{1}{\lambda_2} \right)^{1+\frac{(\alpha-1)\rho}{1-\rho}} \max\left\{ \frac{8\beta+8\lambda_2^2}{\lambda_3^2}, \frac{8+2\lambda_2}{\lambda_1^2} \right\}. \quad (3)$

Then we have

$$E[||\theta_t - \theta^*||_2^2] \leq \left( 1 - \frac{\min\{\lambda_1, \lambda_2\}}{8} \right)^T \Delta_0 + A_1,$$ \quad (4)

where $\Delta_0 = ||w_0 - \omega^*(\theta_0)||_2^2 + ||\theta_0 - \theta^*||_2^2$, where $A_1$ is a constant defined in eq. (3) in Appendix A. Furthermore, let $M \geq \frac{2\lambda_1}{\lambda_2}$ and $T \geq \frac{8}{\min\{\lambda_1, \lambda_2\}} \ln\left( \frac{2a}{e} \right)$. The total sample complexity for Algorithm 1 to achieve an $\epsilon$-accurate optimal solution $\theta^*$, i.e., $E[||\theta_t - \theta^*||_2^2] \leq \epsilon$, is given by

$$TM = \Theta\left( \frac{1}{\epsilon} \log \left( \frac{1}{\epsilon} \right) \right).$$

Theorem 1 shows that the convergence error of Algorithm 1 consists of two terms: the first term is the transient error decreasing at an exponential rate, and the second term is the variance error that diminishes as the batch size $M$ increases. This is in contrast to the single-sample TDC under constant stepsize, which suffers from the variance and bias errors with order $O(\beta/\alpha)$ [Gupta et al., 2019]. Thus, $\epsilon$-level small stepsize $\alpha$ and $\beta$ are required in single-sample TDC to reduce the variance error to achieve the required $\epsilon$-accurate optimal solution, which can slow down the practical convergence speed significantly. In contrast, mini-batch TDC can attain high accuracy with a large constant (not $\epsilon$-level) stepsize. Our result of $O(1/\epsilon \log(1/\epsilon))$ achieves the optimal complexity order due to the lower bound given in [Dalal et al., 2019]. In contrast to the same sample complexity established in [Kaledin et al., 2020], which is applicable only under a diminishing stepsize, our result given in Theorem 1 is applicable under the constant stepsize, which is practically preferred due to the much better performance.

We next provide a sketch of the proof for Theorem 1.

**Proof Sketch of Theorem 1.** The proof of Theorem 1 consists of the following three steps. At $t$-th step, we call $||\theta_t - \theta^*||_2^2$ as the training error and $||w_t - w(\theta_t)||_2^2$ as the tracking error.

**Step 1:** We establish the following induction relationships for the tracking error:

$$E\left[ ||w_{t+1} - w(\theta_{t+1})||_2^2 \right] \leq (1 - \Theta(\lambda_2 + \Theta(\alpha^2/\beta))) E\left[ ||w_t - w(\theta_t)||_2^2 \right] + \Theta(\alpha^2/\beta + \lambda_1 \alpha) E\left[ ||\theta_t - \theta^*||_2^2 \right] + \Theta(1/M). \quad (4)$$

**Step 2:** We then establish the induction relationships for the training error:

$$E\left[ ||\theta_{t+1} - \theta^*||_2^2 \right] \leq (1 - \Theta(\lambda_1 + \Theta(\alpha^2))) E\left[ ||\theta_t - \theta^*||_2^2 \right] + \Theta(\alpha^2 + \lambda_1) E\left[ ||w_t - w(\theta_t)||_2^2 \right] + \Theta(1/M). \quad (5)$$

**Step 3:** Combining eq. (4) and eq. (5) and letting the stepsize $\alpha$ and $\beta$ and batch size $M$ satisfy the requirement specified in Theorem 1, we establish the induction relationship of $\Delta_t = E[||\theta_t - \theta^*||_2^2] + E[||w_t - w(\theta_t)||_2^2]$ as follows:

$$\Delta_{t+1} \leq (1 - \Theta(\min\{\lambda_1, \lambda_2\})) \Delta_t + \Theta(1/M). \quad (6)$$

Applying eq. (6) recursively from $t = T - 1$ to 0 yields the desired convergence result.

4 Two Timescale TDC with Nonlinear Function Approximation

In this section we first introduce the nonlinear two timescale TDC algorithm to solve the policy evaluation problem, then we provide our non-asymptotic convergence rate result.

4.1 Algorithm

In this section we consider policy evaluation problem with nonlinear function approximation, in which a parameterized smooth nonlinear function $v(s, \theta)$ is used to approximate the value function $V^\pi(s)$. [Blatnagar et al., 2009] proposed an algorithm to find a parameter for the approximator $v(s, \theta)$, named nonlinear TDC. The nonlinear TDC updates the parameter by minimizing the following mean-square projected Bellman error objective defined as:

$$J(\theta) = E_{\mu^s}[\hat{v}(s, \theta) - \Pi_\theta T^\top \hat{v}(s, \theta)]^2,$$ \quad (7)

where $\Pi_\theta$ is the orthogonal projection operation into the function space $\mathcal{V} = \{ V(\zeta) \mid \zeta \in \mathbb{R}^d \}$ and $v(s, \zeta) = \phi_\theta(s)^\top \zeta$ with $\phi_\theta(s) = \nabla_{\theta_0} \hat{v}(s, \theta)$. In general, since $J(\theta)$ defined in eq. (7) is nonconvex with respect to the parameter $\theta$, finding the global minimum of $J(\theta)$
Algorithm 2 Two Time-scale Nonlinear TDC

1: Input batch size $M$, learning rate $\alpha$ and $\beta$
2: Sampling: A trajectory $\{s_j, a_j, s_{j+1}\}_{j=0}^{\infty}$ is sampled by following the policy $\pi$
3: Initialization: $\theta_0$ and $w_0$
4: for $t = 0, 1, ..., T$ do
5: $i_t = tM$
6: $w_{t+1} = w_t + \beta \frac{1}{M} \sum_{j=i_t}^{i_t+M-1} (-\phi_{\theta_t}(s_j)\phi_{\theta_t}(s_j)^T w_t + \delta_j(\theta_t)\phi_{\theta_t}(s_j))$
7: $\theta_{t+1} = \theta_t + \alpha \frac{1}{M} \sum_{j=i_t}^{i_t+M-1} (\delta_j(\theta_t)\phi_{\theta_t}(s_j) - \gamma\phi_{\theta_t}(s_{j+1})\phi_{\theta_t}(s_j)^T w_t - h_j(\theta_t, w_t))$
8: end for
Output: $\theta_T$ with $T$ chosen uniformly from $\{1, \ldots, T\}$
Proof Sketch of Theorem 2. The proof of Theorem 2 consists of the following four steps.

Step 1: We first provide Lemma 4 to show that \( w(\theta) \) is \( L_w \)-Lipschitz:

\[
\|w(\theta) - w(\theta')\|_2 \leq L_w \|\theta - \theta'\|_2, \quad \text{for all } \theta, \theta' \in \mathbb{R}^d.
\]

This property is crucial for the convergence analysis of two time-scale nonlinear TDC. It indicates that if \( \theta_t \) changes slowly, then \( w(\theta_t) \) also changes slowly. This allows our finite time analysis to be over a slowly changing linear SA with corresponding ODE defined in eq. 10, guaranteeing that \( \|w_t - w(\theta_t)\|_2^2 \) is small in an amortized sense.

Step 2: We then establish the induction relationships for the tracking error \( \|w_t - w(\theta_t)\|_2^2 \):

\[
\mathbb{E} \left[ \|w_{t+1} - w(\theta_{t+1})\|_2^2 \right] \\
\leq (1 - \Theta(\lambda_0)) \mathbb{E} \left[ \|w_t - w(\theta_t)\|_2^2 \right] \\
+ \Theta(\alpha^2/\beta) \mathbb{E} \left[ \|\nabla J(\theta_t)\|_2^2 \right] + \Theta(1/M). \tag{11}
\]

Step 3: We then establish the induction relationships for the gradient norm \( \|\nabla J(\theta_t)\|_2^2 \):

\[
(\Theta(\alpha) - \Theta(\alpha^2)) \mathbb{E} \left[ \|\nabla J(\theta_t)\|_2^2 \right] \\
\leq \mathbb{E}[J(\theta_t)] - \mathbb{E}[J(\theta_{t+1})] \\
+ \Theta(\alpha + \alpha^2) \mathbb{E} \left[ \|w_t - w(\theta_t)\|_2^2 \right] + \Theta(1/M). \tag{12}
\]

Step 4: Applying eq. 11 and eq. 12 recursively from \( t = T - 1 \) to 0 and combing those two results together yield

\[
(\Theta(\alpha) - \Theta(\alpha^2) - \Theta(\alpha^3)) \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla J(\theta_t)\|_2^2] \\
\leq J(\theta_0) - \mathbb{E}[J(\theta_T)] \\
+ \Theta((\alpha + \alpha^2)/\beta) \|w_0 - w(\theta_0)\|_2^2 + \Theta(1/M).
\]

Letting the stepsize \( \alpha \) and \( \beta \) and the batch size \( M \) satisfies the requirement specified in Theorem 2 we can then obtain the desired convergence result. \( \square \)

5 Policy Optimization: Greedy-GQ Algorithm

In this section, we will provide the non-asymptotic convergence result of Greedy-GQ [Maei et al., 2010], which is also a two timescale nonlinear SA algorithm.

Greedy-GQ was proposed in [Maei et al., 2010] to solve the divergence issue of Q-Learning in the linear function approximation setting. In Greedy-GQ, the goal of the agent is to learn an optimal policy for the MDP with respect to the total expected discounted reward. In the linear function approximation setting, a linear function \( \hat{Q}(s,a,\theta) = \phi(s,a)^\top \theta \) is used to approximate the state-action value function \( Q(s,a) \), where \( \phi(s,a) \in \mathbb{R}^d \) is a fixed feature vector for state-action pair \( (s,a) \) and \( \theta \in \mathbb{R}^d \) is a parameter vector. Without loss of generality, we assume that the feature vector \( \|\phi(s,a)\|_2 \leq 1 \) for all \( (s,a) \in S \times A \) and the columns of the feature matrix \( \Phi \) are linearly independent. In this setting, we hope to find a solution \( \theta \) that satisfies

\[
\Pi T^{\pi_\theta} \hat{Q}(s,a,\theta) = \hat{Q}(s,a,\theta), \quad \text{for all } (s,a) \in S \times A, \tag{13}
\]

where \( \pi_\theta \) is the soft-max greedy policy with respect to the state-action value function \( Q(s,a,\theta) \), i.e., \( \pi_\theta(a|s) = \frac{\exp(\gamma \hat{Q}(s,a,\theta))}{\sum_{a' \in A} \exp(\gamma \hat{Q}(s,a',\theta))} \), where \( \gamma \) is the temperature parameter, and \( T^{\pi_\theta} \) denotes the Bellman operator with policy \( \pi_\theta \). Similarly to the TDC algorithms, Greedy-GQ searches a parameter that satisfies eq. 13 by minimizing a projected Bellman error objective function defined as:

\[
J(\theta) = \mathbb{E}_{\mu_{\pi_\theta}} [\hat{Q}(s,a,\theta) - \Pi T^{\pi_\theta} \hat{Q}(s,a,\theta)]^2, \tag{14}
\]

where \( \delta(\theta) = r(s,a,s') + \gamma \hat{Q}(s',b,\theta) - \hat{Q}(s,a,\theta) \) is the temporal difference error, with \( a \sim \pi_\theta(\cdot|s) \) and \( b \sim \pi_\theta(\cdot|s') \). Since \( J(\theta) \) is nonconvex and smooth everywhere, we can apply gradient descent method to find a local optimal (stationary point) of the objective \( J(\theta) \) via applying the update

\[
\frac{1}{2} \nabla J(\theta_t) = -\mathbb{E}_{\mu_{\pi_\theta}} [\mathbb{E}_{\pi_\theta}[\delta(\theta)|s,a]\phi(s,a)] + \gamma \mathbb{E}_{\mu_{\pi_\theta}} [\mathbb{E}_{\pi_\theta}[\theta(s',b)|s,a]\phi(s,a)^\top w(\theta)]
\]

where

\[
w(\theta) = \mathbb{E}_{\mu_{\pi_\theta}} [\phi(s,a)\phi(s,a)^\top]^{-1} \mathbb{E}_{\mu_{\pi_\theta}} [\mathbb{E}_{\pi_\theta}[\delta(\theta)|s,a]\phi(s,a)].
\]

Similarly to the nonlinear TDC algorithms in section 4, here an auxiliary parameter \( w_t \) is adopted to estimate the value \( \hat{w}(\theta_t) \) by solving a linear SA with the following corresponding ODE:

\[
\dot{w} = -\mathbb{E}_{\mu_{\pi_\theta}} [\phi(s,a)\phi(s,a)^\top] w + \mathbb{E}_{\mu_{\pi_\theta}} [\mathbb{E}_{\pi_\theta}[\delta(\theta)|s,a]\phi(s,a)].
\]

Then, \( \theta_t \) can be updated via direct sampling

\[
\theta_{t+1} = \theta_t + \alpha \frac{1}{|B_t|} \sum_{j \in B_t} g(\theta_t, w_t, x_j),
\]

where \( \rho_0(s,a) = \pi_\theta(a|s)/\pi_\theta(a|s) \) is the importance weighting factor bounded by \( \rho_{\max} \) and we define \( g(\theta_t, w_t, x_j) = \rho_0(s_{j+1}, a_{j+1}) (\delta_j(\theta_t)\phi(s_j, a_j) - \hat{Q}(s_j, a_j, \theta_t)) \).
\(\gamma \phi(s_{j+1}, a_{j+1}) \phi(s_j, a_j)^{\top} w_t.\) The two timescale Greedy-GQ algorithm is shown below.

**Algorithm 3 Two Timescale Greedy-GQ**

1: **Input:** batch size \(M\), learning rate \(\alpha\) and \(\beta\)
2: **Sampling:** A trajectory \(\{s_j, a_j\}_{j \geq 0}\) is sampled by following the behaviour policy \(\pi_b\)
3: **Initialization:** \(\theta_0\) and \(w_0\)
4: for \(t = 0, \cdots, T - 1\) do
5: \(i_t = \mathbb{1} \epsilon M\)
6: \(w_{i_t+1} = w_t + \beta \frac{1}{M} \sum_{j=0}^{i_t} \mathbb{1} \epsilon (\phi(s_j, a_j)\phi(s_j, a_j)^{\top} w_t + \rho_{\theta}(s_j, a_j)\delta_\lambda(\theta_j)(\phi(s_j, a_j)) \gamma \phi(s_{j+1}, a_{j+1}) \phi(s_j, a_j)^{\top} w_t)
7: \(\theta_{i_t+1} = \theta_{i_t} + \alpha \frac{1}{M} \sum_{j=0}^{i_t} \mathbb{1} \epsilon (\phi(s_j, a_j)\phi(s_j, a_j)^{\top} (\phi(s_j)^{\top} w_t - \gamma \phi(s_{j+1}, a_{j+1}) \phi(s_j, a_j)^{\top} w_t))\)
8: end for

Output: \(\theta_T\) with \(\hat{T}\) chosen uniformly from \(\{1, \cdots, T\}\)

By slightly abusing notations in Section 4 we make the follow standard assumptions.

**Assumption 6** (Non-singularity). We have \((\max_{x \in \mathbb{R}^d} \{\lambda_1 \mathbb{E}_{\mu}(A_0^-A_0^\top)\})^{-1} = \lambda_1\) and \(\max_{x \in \mathbb{R}^d} \{C\} = \lambda_2\), where \(A_0 = \mathbb{E}_{\mu_\lambda}(\gamma \mathbb{E}_{\mu_\lambda} \phi(s)^{\top} s - \phi(s)\phi(s)^{\top})\) and \(C = -\mathbb{E}_{\mu_\lambda}(\phi(s)\phi(s)^{\top})\) and \(\lambda_1\) and \(\lambda_2\) are positive constants.

**Assumption 7** (Bounded importance factor). For any state-action pair \((s, a) \in S \times A\) and any \(\theta \in \mathbb{R}^d\), we have \(\rho_\theta(s, a) \leq \rho_{\max}\), where \(\rho_{\max}\) is a positive constant.

Note that Assumption 7 can be satisfied when the behaviour policy is non-degenerated for all states. Moreover, we make the following Lipschitz property of the gradient \(\nabla J(\theta)\), which has been verified in Wang and Zou, 2020 in a similar setting.

**Lemma 1.** Suppose Assumption 7 and Assumption 2 hold, for any \(\theta, \theta' \in \mathbb{R}^d\), we have \(\|\nabla J(\theta) - \nabla J(\theta')\|_2 \leq L_J \|\theta - \theta'\|_2\), where \(L_J\) is a positive constant.

Note that the Greedy-GQ algorithm in Algorithm 3 and nonlinear TDC algorithm in Algorithm 2 share similar structures. Both objectives are nonconvex and both algorithms adopt a two timescale update scheme, in which the fast timescale iteration corresponds to a linear SA and the slow time-scale iteration corresponds to a nonlinear SA. Thus, the analysis of two time-scale nonlinear TDC in Section 4 can be extended to study the convergence rate of Greedy-GQ algorithm. The following theorem characterizes the convergence rate and sample complexity of Algorithm 3.

**Theorem 3.** Consider the two timescale Greedy-GQ algorithm in Algorithm 3. Suppose Assumptions 7 and 2 hold. Let the stepsize \(\beta \leq \min\{\frac{\lambda_1}{2L_J}, \frac{\lambda_2}{8\sqrt{\rho_{\max}}}, \frac{L_J}{3}\}\) and \(\alpha \leq \min\{\frac{\lambda_1}{8L_J}, \frac{\lambda_2}{8\sqrt{\rho_{\max}}}, \frac{L_J}{12\rho_{\max}^2}\}\), and batch size \(M \geq \frac{1 + (\kappa - 1)\rho}{1 - \rho} \max\{128 \left(\rho_{\max}^2 + \frac{1}{N_T}\right) \left[1 + \frac{\lambda_1^2}{8\lambda_2^2} \frac{\lambda_1^2}{\rho_{\max}^2} \frac{\lambda_1^2}{\rho_{\max}^2}\right] + 2\beta^2\} / (\rho_{\max}^2)^2\). We have

\[
\mathbb{E}\left[\left\|\nabla J(\theta_T)\right\|^2_2\right] \leq 8\left(\mathbb{E}[J(\theta_0)] - \mathbb{E}[J(\theta_T)]\right) + \frac{192\rho_{\max}^2 \|w_0 - w^*(\theta_0)\|_2}{\alpha T} + \frac{32C_1 + 1 + (\kappa - 1)\rho}{M (1 - \rho)},
\]

where \(C_1\) is a positive constant defined in eq. (60) in Appendix C. Furthermore, let \(M \geq \frac{64C_2 (1 + (\kappa - 1)\rho)}{(1 - \rho)\epsilon}\) and \(T \geq \frac{8\left(\mathbb{E}[J(\theta_0)] + \frac{192\rho_{\max}^2 \|w_0 - w^*(\theta_0)\|_2}{\alpha \lambda_2^2}\right)}{\epsilon}\). The total sample complexity for Algorithm 3 to achieve an \(\epsilon\)-accurate stationary point, i.e., \(\mathbb{E}\left[\left\|\nabla J(\theta_T)\right\|^2_2\right] \leq \epsilon\), is given by

\[
TM = \Theta\left(\frac{1}{\epsilon^2}\right).
\]

Similarly to Theorem 2 in Theorem 3 we show that Algorithm 3 converges to an \(\epsilon\)-accurate stationary point with sample complexity \(O(\epsilon^{-2})\). Note that Wang and Zou, 2020 studied the convergence rate of two timescale Greedy-GQ with diminishing stepsize, which achieves the complexity of \(O(\epsilon^{-3}\log(1/\epsilon))\). Theorem 3 for two timescale Greedy-GQ with constant stepsize outperforms the result in Wang and Zou, 2020 by a factor of \(O(\epsilon^{-1}\log(1/\epsilon))\), indicating that the constant stepsize can significantly improve the convergence rate of two timescale Greedy-GQ algorithm.

### 6 Conclusion

In this paper, we study the convergence rate for two timescale linear and nonlinear TDC and Greedy-GQ under Markovian sampling and constant stepsize. Specifically, we show that the complexity result of linear TDC orderwisely achieves the optimal convergence rate under a constant stepsize. Our result for nonlinear TDC is the first under Markovian sampling. Moreover, our sample complexity result of Greedy-GQ outperforms the previous result orderwisely. For future work, it is interesting to apply more advance optimization techniques, e.g., acceleration, variance reduction, to further improve the convergence performance of the value-based RL algorithms studied in this paper.

### 7 Acknowledgement

The work was supported partially by the U.S. National Science Foundation under the grants CCF-1801855 and CCF-1900145.
References


Sample Complexity Bounds for Two Timescale Value-based Reinforcement Learning Algorithms


