Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 10, No. 3, 2023

Finite-Time Bounds for AMSGrad-Enhanced Neural TD

Tiange Fu, Qingtao Wu

School of Information Engineering, Henan University of Science and Technology, Luoyang, 471023, China

Abstract: Although the combination of adaptive methods and deep reinforcement learning has achieved tremendous success in practical applications, its theoretical convergence properties are not well understood. To address this issue, we propose a neural network-based adaptive TD algorithm, called NTD-AMSGrad, which is a variant of temporal difference learning. Moreover, we rigorously analyze the convergence performance of the proposed algorithm and establish a finite-time bound for NTD-AMSGrad under the Markov observation model. Specifically, when the neural network is wide enough, the proposed algorithm can converge to the optimal action-value function at a rate of, where is the number of iterations.

Keywords: Reinforcement learning; Temporal difference learning; Neural networks; Adaptive methods.

1. Introduction

Reinforcement learning (RL) has garnered considerable attention in recent years due to its wide-ranging applications, including medical diagnosis [1], financial quantization [2], chat generative pre-trained transformer [3], smart grid [4], and many more. At its core, RL involves the agent interacting with the environment in a trial-and-error process to learn an optimal policy that maximizes its cumulative long-term reward [5]. A key challenge in developing reinforcement learning algorithms is estimating the long-term reward associated with a given policy. This problem, which is commonly referred to as policy evaluation, is of fundamental importance in reinforcement learning.

Temporal-difference learning (TD), originally proposed by Sutton [6], is a crucial approach for solving the policy evaluation problem by estimating the value of a state or action based on the difference between the predicted and actual reward received at each time step. TD has been demonstrated to be both efficient and effective in a wide range of tasks, including e-sports games and robot control [7, 8]. In addition, TD can be easily combined with function approximation, such as neural networks, to learn effective policies in high-dimensional state spaces.

Despite its many advantages, TD still faces several challenges, particularly with respect to theoretical analysis in the context of nonlinear function approximators. Nonlinear approximators can introduce issues such as instability and divergence in TD algorithms [9, 10, 11], which can make it difficult to learn an accurate value function. As a result, developing effective techniques for TD with nonlinear function approximation remains an active area of research.

In this context, recent work has explored the use of neural networks as function approximators in TD [7, 12, 13]. These approaches have shown significant promise in addressing the challenges associated with nonlinear function approximation and have led to breakthrough results in a variety of applications. Furthermore, researchers in [14, 15, 16] have presented convergence results for neural TD, albeit under certain additional assumptions and restrictions. In an effort to enhance the efficiency of TD, adaptive methods inspired by stochastic algorithms have been proposed in [17, 18] for Deep Q-Networks (DQN). Empirical evidence indicates that these adaptive TD variants outperform their vanilla counterparts in numerous tasks. However, there is still much to be learned about the theoretical properties of these algorithms and the conditions under which they are guaranteed to converge to an optimal policy. As such, further research is needed to fully understand the potential of neural TD learning and to develop stable algorithms that can be applied to a wide range of real-world problems.

To address this research gap, the present study proposes an Adam-type TD algorithm with neural network approximation, termed as NTD-AMSGrad. The proposed algorithm combines AMSGrad algorithm [19] with neural TD and adaptively adjusts the learning rate of different weights and biases in the neural network by utilizing the moving average of historical gradients. Moreover, we provide a rigorous non-asymptotic convergence analysis of the proposed algorithm under Markov observation. Additionally, we elaborate on the key contributions of this paper, which are summarized below:

We propose an adaptive TD with neural network approximation called NTD-AMSGrad under Markovian sampling

We demonstrate that, given a sufficiently wide neural network, NTD-AMSGrad can converge to the optimal action-value function at a rate of $\mathcal{O}(1/\sqrt{T})$, where T is the number of iterations.

The remainder of this paper is organized as follows. In Section 2, we introduce the necessary preliminaries. In Section 3, we formulate reinforcement learning problem with neural network function approximation. To solve this problem, we propose NTD-AMSGrad and provide the standard assumptions. The main results of this paper are presented in Section 4. In Section 5, we provide the rigorous proofs of main results in detail. Finally, we conclude this paper in Section 6.

The subsequent sections of this paper are structured as follows. Section 2 presents the necessary preliminaries. In Section 3, we formulate the reinforcement learning problem with neural network function approximation. To address this problem, we propose NTD-AMSGrad and outline the standard assumptions. In Section 4, we provide detailed and rigorous proofs of the main theorems. Finally, Section 5 concludes this paper.

2. Preliminaries

In this section, we provide the necessary preliminaries for the policy evaluation problem. We denote the set of states as

$\mathcal{S}$. Similarly, the set of actions is denoted as $\mathcal{A}$. The Markovian transition probability matrix is represented by $\mathcal{P}$, and the reward function is denoted as $\mathcal{R}$. Therefore, a Markov reward process can be defined using a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\gamma \in (0,1)$ represents the discount factor. For any given policy π, the associated value function $V_\pi : \mathcal{S} \to \mathbb{R}$ is denoted as

$$V_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s \right],$$

the corresponding action-value function can be defined as

$$Q_\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a \right].$$

The properties of Markov reward process leads to the Bellman equation, which is given by

$$\mathcal{T} Q_\pi(s,a) = Q_\pi(s,a), \tag{1}$$

where $\mathcal{T}$ is the Bellman operator, and Q^* is the unique fixed point of $\mathcal{T}$.

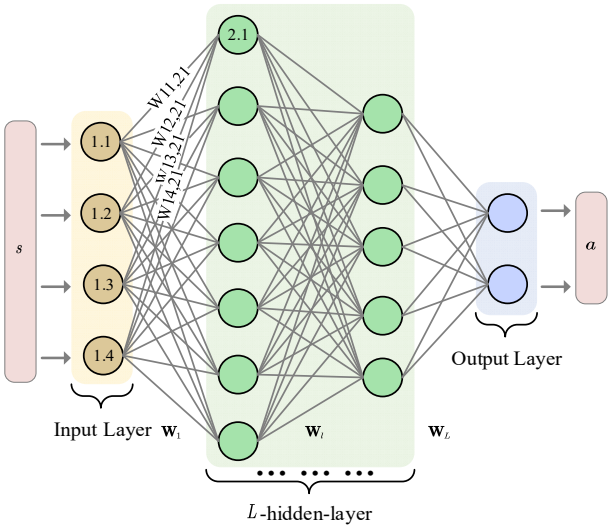

Fig. 1 ReLU network diagram

In this paper, we consider the L-hidden-layer ReLU network to approximate the action-value function Q, as depicted in Figure 1. The formulation is as follows

$$f(\theta; \mathbf{x}) = \sqrt{m} \cdot \mathbf{W}_L \sigma(\mathbf{W}_{L-1} \cdots \sigma(\mathbf{W}_1 \mathbf{x}) \cdots), \tag{2}$$

where the input parameter is represented by the vector $\mathbf{x} \in \mathbb{R}^d$, the variables m and L correspond to the width and depth of the neural network, respectively. For $l = 2, \ldots, L-1$, the matrices $\mathbf{W}_1 \in \mathbb{R}^{m \times d}$, $\mathbf{W}_L \in \mathbb{R}^{1 \times m}$, $\mathbf{W}_l \in \mathbb{R}^{m \times m}$ are involved. Additionally, all parameter matrices are vectorized as $\theta = (\mathrm{vec}(\mathbf{W}_1)^\top, \ldots, \mathrm{vec}(\mathbf{W}_L)^\top)^\top$. Thus, the action-value function approximated by neural network can be expressed as

$$Q_\pi(s,a) \approx f(\theta; \phi(s,a)), \tag{3}$$

where $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a feature mapping?

3. Problem Setup, Algorithm Design, and Assumptions

Assumption 1. For any state-action pair, we make the assumption that the feature vector is uniformly bounded, i.e.,

Assumption 2. Suppose that the Markov chain $\mathcal{P}$ is irreducible.

Algorithm 1 NTD-AMSGrad

Input: Parameters β_1, β_2, α_t, π, γ, $\mathbf{W}_l(0) \sim \mathcal{N}(0, 1/m)$.

Output: θ_t

1: **Initialization:** $\theta_0 = (\mathbf{W}_1(0)^\top, \ldots, \mathbf{W}_L(0)^\top)^\top$, $m_0 = 0$, $v_0 = 0$

2: **for** $t = 0, 1, \ldots$ **do**

3: $g_t = \nabla_\theta f(\theta_t; \phi(s_t, a_t)) \Delta_t$

4: $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$

5: $v_t = \beta_2 v_{t-1} + (1-\beta_2)\, g_t \odot g_t$

6: $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$

7: $\theta_{t+1} = \Pi_{\mathcal{X}, \sqrt{\hat{v}_t}} \left(\theta_t - \alpha_t \frac{m_t}{\sqrt{\hat{v}_t}} \right)$

8: **end for**

Define the locally linearized space $\mathcal{F} := \{ f(\theta_0; \phi(s,a)) + \langle \nabla_\theta f(\theta_0, \phi(s,a)), \theta - \theta_0 \rangle : \theta \in \Theta \}$, where Θ represents a constraint set. Then, solving the policy evaluation problem transforms into minimizing the mean square projected Bellman error (MSPBE), as expressed in the following equation

$$\min \mathbb{E}_{\mu,\pi,\mathcal{P}} \left[(Q(s,a;\theta) - \Pi_{\mathcal{F}} \mathcal{T} Q(s,a;\theta))^2 \right], \tag{4}$$

where $\Pi_{\mathcal{F}}$ denotes the projection operator. Subsequently, we outline the neural TD approach. The updates in neural TD are carried out using the gradient, where the gradient term is defined as follows

$$\begin{aligned} g(\theta; s_t, a_t, s_{t+1}, a_{t+1}) &= \nabla_\theta f(\theta; \phi(s_t, a_t)) \left(f(\theta; \phi(s_t, a_t)) - r_t - \gamma f(\theta; \phi(s_{t+1}, a_{t+1})) \right) \\ &= \nabla_\theta f(\theta; \phi(s_t, a_t)) \Delta_t(s_t, a_t, s_{t+1}; \theta), \end{aligned} \tag{5}$$

where Δ_t represents the TD error.

Definition 1. [20] A point $\theta^* \in \Theta$ is said to be the approximate stationary if

$$\mathbb{E}_{\mu,\pi,\mathcal{P}} \left[\hat{\Delta}_t \left\langle \nabla_\theta \hat{f}(\theta^*; \phi(s,a)), \theta - \theta^* \right\rangle \right] \geq 0, \tag{6}$$

where $\hat{f}(\theta; \phi(s,a)) \in \mathcal{F}$ and the TD error is

$$\hat{\Delta}_t(s, a, s', a'; \theta) = \hat{f}(\theta; \phi(s,a)) - r_t - \gamma \hat{f}(\theta; \phi(s', a')). \tag{7}$$

Cai et al. [20] have proved the existence of an approximate stationary point that minimizes the MSPBE. In addition, we introduce a vector-value map gradient that remains independent of the data point and is defined as

$$\overline{g}(\theta; s_t, a_t, s_{t+1}, a_{t+1}) = \mathbb{E}_{\mu,\pi,\mathcal{P}} \left[\nabla_\theta f(\theta; \phi(s_t, a_t)) \hat{\Delta}_t(s_t, a_t, s_{t+1}; \theta) \right]. \tag{8}$$

Likewise, based on the linearized function, the following gradient terms is given by

$$h(\theta) = \hat{\Delta}(s_t, a_t, s_{t+1}, a_{t+1}; \theta) \nabla_\theta \hat{f}(\theta; \phi(s_t, a_t)), \tag{9}$$

$$\overline{h}(\theta)=\mathbb{E}_{\mu,\pi,\mathcal{P}}\left[\hat{\Delta}(s,a,,s',a';\theta)\nabla_\theta\hat{f}(\theta;\phi(s_t,a_t))\right]. \tag{10}$$

Next, in conjunction with the AMSGrad, we introduce the remaining definitions of Algorithm 1. The first order moment m_t is updated by the following rule,

$$m_t=\beta_1 m_{t-1}+(1-\beta_1)g_t, \tag{11}$$

where $0\leq\beta_1<1$ is a hyper-parameter? Furthermore, the second order moment v_t is defined as

$$v_t=\beta_2 v_{t-1}+(1-\beta_2)\,g_t\odot g_t, \tag{12}$$

where $0\leq\beta_2<1$ denotes a hyper-parameter, and $\hat{v}_t=\max\left(\hat{v}_{t-1},v_t\right)$. Then, the parameter θ_t is updated as

$$\theta_{t+1}=\Pi_{\mathcal{X},\sqrt{\hat{v}_t}}\left(\theta_t-\alpha_t\frac{m_t}{\sqrt{\hat{v}_t}}\right), \tag{13}$$

where $\Pi_{\mathcal{X},\sqrt{\hat{v}_t}}(\cdot)$ is a weighted projection operator onto $\mathcal{X}$, which is given by

$$\Pi_{\mathcal{X},v}(a)=\arg\min_{a\in\mathcal{X}}\|a-b\|_v^2. \tag{14}$$

Thus, this section devises an adaptive TD with neural network approximation, named NTD-AMSGrad, which is summarized in Algorithm 1. Before delving into the analysis of the finite-time bounds of NTD-AMSGrad, it is imperative to introduce the following set of standard assumptions.

4. Convergence Analysis

In this section, we will rigorously analyze the convergence performance of NTD-AMSGrad. To accomplish this, we will present several key results that are essential for our analysis.

Lemma 1. For $\forall\theta\in\Theta$, the following relations holds

$$\langle\overline{h}(\theta)-\overline{h}(\theta^*),\theta-\theta^*\rangle\geq(1-\zeta^{-\frac{1}{2}})\mathbb{E}\left[(\hat{f}(\theta)-\hat{f}(\theta^*))^2|\theta_0\right].$$

Theorem 1. Let Assumptions 1 and 2 hold. For $t\geq0$, the sequences $\{\theta_t\}$, $\{m_t\}$ and $\{v_t\}$ are generated by Algorithm 1. Moreover, let $0\leq\beta_1<1$, $0\leq\beta_2<1$, $\delta=\frac{\beta_1^2}{\beta_2}<1$, the width of ReLU network $m\geq C_0\max\{dL^2\log(m/\omega),\sigma^{-\frac{4}{3}}L^{-\frac{8}{3}}\log(m/(\sigma\omega))\}$, and the radius $\psi=C_1m^{-\frac{1}{2}}L^{-\frac{9}{4}}$. Then, with the probability at least $1-2\omega-L^2\exp(-C_2m^{\frac{2}{3}}L)$ over θ_0, the following bound holds

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\hat{f}\left(\theta_t\right)-\hat{f}\left(\theta^*\right)\right)^2\mid\theta_0\right]\leq&\frac{D^2\sqrt{T}}{2\alpha\left(1-\beta_1\right)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2}+\frac{\alpha\sqrt{1+\log T}}{\sqrt{\left(1-\beta_2\right)\left(1-\delta\right)}}\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}g_{t,i}^2}\\&+\frac{C_2\tau^*\log(T/\omega)\log T}{\lambda\sqrt{T}}+\frac{C_3\sqrt{\log m\log(T/\omega)}}{\lambda m^{1/6}},\end{aligned}$$

where $\lambda=1-\zeta^{-\frac{1}{2}}\in(0,1)$, τ^* is the mixing time of the Markov chain, and $\{C_i>0\}_{i=0,\ldots,3}$ represents the universal constants.

Theorem 2. Under the same conditions of theorem 1, NTD-AMSGrad yields the following bound with probability at least $1-3\omega-L^2\exp(-C_0m^{\frac{2}{3}}L)$ over the randomness of θ_0

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(Q\left(s,a;\theta_t\right)-Q^*(s,a)\right)^2\right]\leq&\frac{3\mathbb{E}\left[\left(\Pi Q^*(s,a)-Q^*(s,a)\right)^2\right]}{(1-\gamma)^2}+\frac{D^2\sqrt{T}}{2\alpha\left(1-\beta_1\right)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2}\\&+\frac{\alpha\sqrt{1+\log T}}{\sqrt{\left(1-\beta_2\right)\left(1-\delta\right)}}\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}g_{t,i}^2}\\&+\frac{C_1\tau^*\log(T/\omega)\log T}{\lambda\sqrt{T}}+\frac{C_2\sqrt{\log(T/\omega)\log m}}{\lambda m^{1/6}},\end{aligned}$$

where $\{C_i>0\}_{i=0,\ldots,2}$ are the positive constants.

Proof of Theorem 1. Using the polynomial addition and subtraction rules, $\left\langle\overline{h}(\theta_t),\theta_t-\theta^*\right\rangle$ can be decomposed into

$$\left\langle\overline{h}(\theta_t),\theta_t-\theta^*\right\rangle=\left\langle\overline{h}\left(\theta_t\right)-h_t\left(\theta_t\right),\theta_t-\theta^*\right\rangle+\left\langle h_t\left(\theta_t\right)-g_t\left(\theta_t\right),\theta_t-\theta^*\right\rangle+\left\langle g_t\left(\theta_t\right),\theta_t-\theta^*\right\rangle. \tag{15}$$

Furthermore, according to the definition of m_t in Eq. (11), we get

$$\begin{aligned}\left\langle g_t\left(\theta_t\right),\theta_t-\theta^*\right\rangle=&\left\langle\frac{m_t}{1-\beta_1}-\frac{\beta_1m_{t-1}}{1-\beta_1},\theta_t-\theta^*\right\rangle\\=&\frac{1}{1-\beta_1}\left\langle m_t,\theta_t-\theta^*\right\rangle-\frac{\beta_1}{1-\beta_1}\left\langle m_{t-1},\theta_t-\theta^*\right\rangle\\=&\frac{1}{1-\beta_1}\left\langle m_t,\theta_t-\theta^*\right\rangle-\frac{\beta_1}{1-\beta_1}\left\langle m_{t-1},\theta_{t-1}-\theta^*\right\rangle-\frac{\beta_1}{1-\beta_1}\left\langle m_{t-1},\theta_t-\theta_{t-1}\right\rangle\\=&\frac{1}{1-\beta_1}\left\langle m_t,\theta_t-\theta^*\right\rangle-\frac{1}{1-\beta_1}\left\langle m_{t-1},\theta_{t-1}-\theta^*\right\rangle+\left\langle m_{t-1},\theta_{t-1}-\theta^*\right\rangle\\&-\frac{\beta_1}{1-\beta_1}\left\langle m_{t-1},\theta_t-\theta_{t-1}\right\rangle.\end{aligned} \tag{16}$$

Plugging Eq. (16) into Eq. (15), we obtain

$$\begin{aligned}\left\langle \bar{h}\left(\theta_t\right), \theta_t-\theta^*\right\rangle &= \left\langle \bar{h}\left(\theta_t\right)-h_t\left(\theta_t\right), \theta_t-\theta^*\right\rangle + \left\langle h_t\left(\theta_t\right)-g_t\left(\theta_t\right), \theta_t-\theta^*\right\rangle \\ &+ \frac{1}{1-\beta_1}\left\langle m_t, \theta_t-\theta^*\right\rangle - \frac{1}{1-\beta_1}\left\langle m_{t-1}, \theta_{t-1}-\theta^*\right\rangle \\ &+ \left\langle m_{t-1}, \theta_{t-1}-\theta^*\right\rangle - \frac{\beta_1}{1-\beta_1}\left\langle m_{t-1}, \theta_t-\theta_{t-1}\right\rangle .\end{aligned} \tag{17}$$

For $\forall \theta \in \mathcal{X}$, $\left\langle \bar{h}\left(\theta^*\right), \theta-\theta^*\right\rangle \geq 0$, then we get

$$\left\langle \bar{h}\left(\theta_t\right)-\bar{h}\left(\theta^*\right), \theta_t-\theta^*\right\rangle \leq \left\langle \bar{h}\left(\theta_t\right), \theta_t-\theta^*\right\rangle . \tag{18}$$

Based on Lemma 1 and utilizing Eq. (18), we obtain the following

$$\begin{aligned}\mathbb{E}\left[\hat{f}\left(\theta_t\right)-\hat{f}\left(\theta^*\right) \mid \theta_0\right] &\leq \frac{1}{\lambda}\left\langle \bar{h}\left(\theta_t\right)-\bar{h}\left(\theta^*\right), \theta_t-\theta^*\right\rangle \\ &\leq \frac{1}{\lambda}\left\langle \bar{h}\left(\theta_t\right), \theta_t-\theta^*\right\rangle .\end{aligned} \tag{19}$$

Then, substituting Eq. (17) into Eq. (19) yields

$$\begin{aligned}\mathbb{E}\left[\hat{f}\left(\theta_t\right)-\hat{f}\left(\theta^*\right) \mid \theta_0\right] &\leq \left\langle \bar{h}\left(\theta_t\right)-h_t\left(\theta_t\right), \theta_t-\theta^*\right\rangle + \left\langle h_t\left(\theta_t\right)-g_t\left(\theta_t\right), \theta_t-\theta^*\right\rangle \\ &+ \frac{1}{1-\beta_1}\left\langle m_t, \theta_t-\theta^*\right\rangle - \frac{1}{1-\beta_1}\left\langle m_{t-1}, \theta_{t-1}-\theta^*\right\rangle \\ &+ \left\langle m_{t-1}, \theta_{t-1}-\theta^*\right\rangle - \frac{\beta_1}{1-\beta_1}\left\langle m_{t-1}, \theta_t-\theta_{t-1}\right\rangle .\end{aligned} \tag{20}$$

By summing the above inequality over $t=1,\ldots,T$ and utilizing the fact that $m_0=0$, we obtain

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\hat{f}\left(\theta_t\right)-\hat{f}\left(\theta^*\right) \mid \theta_0\right] &\leq \sum_{t=1}^{T}\left\langle \bar{h}\left(\theta_t\right)-h_t\left(\theta_t\right), \theta_t-\theta^*\right\rangle + \sum_{t=1}^{T}\left\langle h_t\left(\theta_t\right)-g_t\left(\theta_t\right), \theta_t-\right. \\ &+ \frac{1}{1-\beta_1}\left(\left\langle m_t, \theta_t-\theta^*\right\rangle - \left\langle m_0-\theta_0-\theta^*\right\rangle\right) + \left\langle m_0, \theta_0-\theta^*\right\rangle \\ &+ \sum_{t=1}^{T-1}\left\langle m_t, \theta_t-\theta^*\right\rangle + \frac{\beta_1}{1-\beta_1}\sum_{t=1}^{T}\left\langle m_{t-1}, \theta_{t-1}-\theta_t\right\rangle \\ &\leq \underbrace{\sum_{t=1}^{T}\left\langle \bar{h}(\theta_t)-h_t(\theta_t), \theta_t-\theta^*\right\rangle}_{I_1} + \underbrace{\sum_{t=1}^{T}\left\langle h_t(\theta_t)-g_t(\theta_t), \theta_t-\theta^*\right\rangle}_{I_2} \\ &+ \frac{\beta_1}{1-\beta_1}\underbrace{\left\langle m_t, \theta_t-\theta^*\right\rangle}_{I_3} + \underbrace{\sum_{t=1}^{T}\left\langle m_t, \theta_t-\theta^*\right\rangle}_{I_4} \\ &+ \frac{\beta_1}{1-\beta_1}\underbrace{\sum_{t=1}^{T}\left\langle m_{t-1}, \theta_{t-1}-\theta_t\right\rangle}_{I_5} .\end{aligned} \tag{21}$$

We will individually bound each term on the right-hand side of Eq. (21) and subsequently combine these bounds. Firstly, following the results of [16], we can obtain bounds for the terms I_1 and I_2

$$I_1 \leq C_0\left(m\log(T/\omega)+m^2\sigma^2\right)\tau^*\rho_{\max\{0,t-\tau^*\}}, \tag{22}$$

where τ^* is the mixing time of the Markov chain, and the term I_2 is

$$I_2 \leq C_1(2+\gamma)m^{-\frac{1}{6}}\sqrt{\log m\log(T/\omega)}, \tag{23}$$

with probability at least $1-2\omega-3L^2\exp(-C_1 m\sigma^{\frac{2}{3}}L)$. Now, we bound for the term I_4. By the nonexpansiveness property of Eq. (14), we obtain

$$\begin{aligned}\left\|\theta_{t+1}-\theta^*\right\|^2_{\hat{v}_t^{1/2}} &= \left\|\Pi_{\mathcal{X},\sqrt{\hat{v}_t}}\left(\theta_t-\alpha_t\sqrt{\hat{v}_t}m_t\right)-\theta^*\right\|^2_{\hat{v}_t^{1/2}} \\ &\leq \left\|\theta_t-\alpha_t\sqrt{\hat{v}_t}m_t-\theta^*\right\|^2_{\hat{v}_t^{1/2}} \\ &= \left\|\theta_t-\theta^*\right\|^2_{\hat{v}_t^{1/2}} - 2\alpha_t\left\langle m_t, \theta_t-\theta^*\right\rangle + \left\|\alpha_t\sqrt{\hat{v}_t}m_t\right\|^2_{\hat{v}_t^{1/2}} \\ &= \left\|\theta_t-\theta^*\right\|^2_{\hat{v}_t^{1/2}} - 2\alpha_t\left\langle m_t, \theta_t-\theta^*\right\rangle + \alpha_t^2\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}} .\end{aligned} \tag{24}$$

Then, dividing both sides of Eq. (24) by $2\alpha_t$ to get

$$\begin{aligned}\left\langle m_t, \theta_t-\theta\right\rangle &\leq \frac{1}{2\alpha_t}\left\|\theta_t-\theta^*\right\|^2_{\hat{v}_t^{1/2}} - \frac{1}{2\alpha_t}\left\|\theta_{t+1}-\theta^*\right\|^2_{\hat{v}_t^{1/2}} + \frac{\alpha_t}{2}\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}} \\ &= \frac{1}{2\alpha_{t-1}}\left\|\theta_\theta-\theta^*\right\|^2_{\hat{v}_{t-1}^{1/2}} - \frac{1}{2\alpha_t}\left\|\theta_{\theta+1}-\theta^*\right\|^2_{\hat{v}_t^{1/2}} \\ &+ \frac{1}{2}\sum_{i=1}^{d}\left(\frac{\hat{v}_{t,i}^{1/2}}{\alpha_t}-\frac{\hat{v}_{t-1,i}^{1/2}}{\alpha_{t-1}}\right)\left(\theta_{t,i}-\theta_i\right)^2 + \frac{\alpha_t}{2}\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}} \\ &\leq \frac{1}{2\alpha_{t-1}}\left\|\theta_t-\theta^*\right\|^2_{\hat{v}_{t-1}^{1/2}} - \frac{1}{2\alpha_t}\left\|\theta_{t+1}-\theta^*\right\|^2_{\hat{v}_t^{1/2}} \\ &+ \frac{D^2}{2}\sum_{i=1}^{d}\left(\frac{\hat{v}_{t,i}^{1/2}}{\alpha_t}-\frac{\hat{v}_{t-1,i}^{1/2}}{\alpha_{t-1}}\right) + \frac{\alpha_t}{2}\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}},\end{aligned} \tag{25}$$

where we use the fact that $\hat{v}_{t,i} \geq \hat{v}_{t-1,i}$, $\frac{1}{\alpha_t} \geq \frac{1}{\alpha_{t-1}}$, and the assumption $D=\max_{a,b\in\mathcal{X}}\left\|a-b\right\|_\infty$. Summing Eq. (25) over $t=1,\ldots,T$, we have

$$I_4 = \sum_{t=1}^{T}\left\langle m_t, \theta_t-\theta^*\right\rangle \leq \frac{D^2}{2\alpha_T}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \frac{1}{2}\sum_{t=1}^{T}\alpha_t\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}}, \tag{26}$$

where $\frac{1}{2\alpha_0}\left\|\theta_1-\theta\right\|^2_{\hat{v}_0^{1/2}}=0$. Further, According to Young inequality and nonexpansiveness of projection, let us bound the term I_5

$$\begin{aligned}I_5 &= \sum_{t=2}^{T}\left\langle m_{t-1}, \theta_{t-1}-\theta_t\right\rangle = \sum_{t=1}^{T-1}\left\langle m_t, \theta_t-\theta_{t+1}\right\rangle \\ &\leq \sum_{t=1}^{T-1}\left\|m_t\right\|_{\hat{v}_t^{-1/2}}\left\|\theta_{t+1}-\theta_t\right\|_{\hat{v}_t^{1/2}} \\ &= \sum_{t=1}^{T-1}\left\|m_t\right\|_{\hat{v}_t^{-1/2}}\left\|\Pi_{\mathcal{X},\sqrt{\hat{v}_t}}\left(\theta_t-\alpha_t\hat{v}_t^{-1/2}m_t\right)-\Pi_{\mathcal{X},\sqrt{\hat{v}_t}}\left(\theta_t\right)\right\|_{\hat{v}_t^{1/2}} \\ &\leq \sum_{t=1}^{T-1}\alpha_t\left\|m_t\right\|_{\hat{v}_t^{-1/2}}\left\|\hat{v}_t^{-1/2}m_t\right\|_{\hat{v}_t^{1/2}} = \sum_{t=1}^{T-1}\alpha_t\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}} .\end{aligned} \tag{27}$$

In addition, using the property of Holder and Young inequality, the bound of the term I_3 can be constrained as

$$\begin{aligned}I_3 &\leq \left\|m_T\right\|_{\hat{v}_T^{-1/2}}\left\|\theta_T-\theta^*\right\|_{\hat{v}_T^{1/2}} \\ &\leq \alpha_T\left\|m_T\right\|^2_{\hat{v}_T^{-1/2}} + \frac{1}{4\alpha_T}\left\|\theta_T-\theta^*\right\|^2_{\hat{v}_T^{1/2}} \\ &\leq \alpha_T\left\|m_T\right\|^2_{\hat{v}_T^{-1/2}} + \frac{D^2}{4\alpha_T}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} .\end{aligned} \tag{28}$$

Now, combining the bounds of terms I_3, I_4 and I_5 yields

$$\begin{aligned}&I_3+I_4+I_5 \\ &= \frac{\beta_1}{1-\beta_1}\left(\left\langle m_T, \theta_T-\theta^*\right\rangle + \sum_{t=1}^{T}\left\langle m_{t-1}, \theta_{t-1}-\theta_t\right\rangle\right) + \sum_{t=1}^{T}\left\langle m_t, \theta_t-\theta^*\right\rangle \\ &\leq \frac{\beta_1}{1-\beta_1}\left(\frac{D^2}{4\alpha_T}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \sum_{t=1}^{T}\alpha_t\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}}\right) + \frac{D^2}{2\alpha_T}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \frac{1}{2}\sum_{t=1}^{T}\alpha_t\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}} \\ &= \frac{(2-\beta_1)D^2}{4\alpha_T(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \frac{1+\beta_1}{2(1-\beta_1)}\sum_{t=1}^{T}\alpha_t\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}} \\ &\leq \frac{D^2\sqrt{T}}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \frac{1}{1-\beta_1}\sum_{t=1}^{T}\alpha_t\left\|m_t\right\|^2_{\hat{v}_t^{-1/2}} \\ &\leq \frac{D^2\sqrt{T}}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \frac{\alpha\sqrt{1+\log T}}{\sqrt{(1-\beta_2)(1-\delta)}}\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}g_{t,i}^2},\end{aligned} \tag{29}$$

where the second inequality uses the assumption $\frac{2-\beta_1}{4} \leq \frac{1}{2}$, $\frac{1+\beta_1}{2} \leq 1$, and $\alpha_T = \frac{\alpha}{\sqrt{T}}$. Finally, with with the probability at least $1-2\omega-L^2\exp(-C_5 m^{\frac{2}{3}}L)$

over the randomness of θ_0, and by combining the bounds of terms I_1 and I_2, we can derive Theorem 1 as follows

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\hat{f}\left(\theta_t\right)-\hat{f}\left(\theta^*\right)\right)^2 \mid \theta_0\right] \\
&\leq \frac{D^2\sqrt{T}}{2\alpha\left(1-\beta_1\right)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2}+\frac{\alpha\sqrt{1+\log T}}{\sqrt{\left(1-\beta_2\right)\left(1-\delta\right)}}\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}g_{t,i}^2} \\
&\quad+\frac{C_0(2+\gamma)m^{-1/6}\sqrt{\log m\log(T/\omega)}}{\lambda} \\
&\quad+\frac{C_1(2+\gamma)^2\log(T/\omega)\log T}{\lambda\sqrt{T}}+\frac{C_2(\log(T/\omega)+1)\tau^*\log T}{\lambda\sqrt{T}} \\
&\leq \frac{D^2\sqrt{T}}{2\alpha\left(1-\beta_1\right)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2}+\frac{\alpha\sqrt{1+\log T}}{\sqrt{\left(1-\beta_2\right)\left(1-\delta\right)}}\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}g_{t,i}^2} \\
&\quad+\frac{C_3\tau^*\log(T/\omega)\log T}{\lambda\sqrt{T}}+\frac{C_4\sqrt{\log m\log(T/\omega)}}{\lambda m^{1/6}},
\end{aligned} \tag{30}$$

where we use the fact that $\gamma<1$ and $\left\|\theta_0-\theta^*\right\|_2^2\leq\frac{1}{m}$. To this end, the result of Theorem 1 is dereived.

Proof of Theorem 2. By the proof approach in [16], we can establish the result of Theorem 2.

According to the Theorem 1, the theoretical analysis demonstrates that NTD-AMSGrad can converge to the minimizer of MSPBE at a rate of $\tilde{O}(1/\sqrt{K})$ when the width of the ReLU neural networks is sufficiently large. Moreover, Theorem 2 presents the finite-time bounds for NTD-AMSGrad.

5. Conclusion

This paper presents a novel neural TD algorithm named NTD-AMSGrad, which adopts the Adam-type optimization method. The finite-time bound of the proposed algorithm is established under Markov observation. Specifically, the theoretical results demonstrate that the proposed algorithm achieves convergence at a rate of $\mathcal{O}(1/\sqrt{T})$, when the neural network is sufficiently wide.

References

[1] N. Salpea, P. Tzouveli, D. Kollias. Medical image segmentation: A review of modern architec-tures. Computer Vision–ECCV 2022 Workshops. Expo Tel Aviv, 2022, pp. 691-708.

[2] N. Lang, N. Shlezinger. Joint privacy enhancement and quantization in federated learning. IEEE Transactions on Signal Processing. Vol. 71 (2023), pp. 295-310.

[3] T. Brown, B. Mann, N. Ryder, et al. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual, 2020, pp. 1877-1901.

[4] A. Kumari, S. Tanwar. A reinforcement-learning-based secure demand response scheme for smart grid system. IEEE Internet of Things Journal. Vol. 9 (2022) No. 3, pp. 2180-2191.

[5] S. R Sutton, G. A Barto. Reinforcement learning: An introduction. 2nd ed. MA: MIT press, 2018, pp. 1-552.

[6] S. R Sutton. Learning to predict by the methods of temporal differences. Machine Learning. Vol. 3 (1988) No. 1, pp. 9-44.

[7] D. Ye, Z. Liu Z, M. Sun M, et al. Mastering complex control in moba games with deep reinforcement learning. Proceedings of the 34th AAAI Conference on Artificial Intelligence. NY, USA, 2020, pp. 6672-6679.

[8] M. Dalal, D. Pathak, R. R Salakhutdinov. Accelerating robotic reinforcement learning via pa-rameterized action primitives. Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual, 2021, pp. 21847-21859.

[9] S. R Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. Proceedings of the 9th International Conference on Neural Information Processing Systems. CO, USA, 1995, pp. 1038-1044.

[10] J. Tsitsiklis, B. Roy Van. Analysis of temporal-diffference learning with function approxima-tion. Proceedings of the 10th International Conference on Neural Information Processing Systems. CO, USA, 1996, pp. 1075-1081.

[11] H. Maei, C. Szepesvari, S. Bhatnagar, et al. Convergent temporal-difference learning with arbi-trary smooth function approximation. Proceedings of the 23rd International Conference on Neural Information Processing Systems. British Columbia, Canada, 2009, pp. 1204-1212.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning. Nature. Vol. 518 (2015) No. 7540, pp. 529-533.

[13] J. Zou, T. Hao, C. Yu, et al. A3C-DO: A regional resource scheduling framework based on deep reinforcement learning in edge scenario. IEEE Transactions on Computers. Vol. 70 (2020) No. 2, pp. 228-239.

[14] Q. Cai, Z. Yang, D. J Lee, et al. Neural temporal-difference learning converges to global opti-ma. Proceedings of the 33rd International Conference on Neural Information Processing Systems. BC, Canada, 2019, pp. 11312-11322.

[15] J. Fan, Z. Wang, Y. Xie, et al. A theoretical analysis of deep q-learning Proceedings of the 2nd Conference on Learning for Dynamics and Control. Berkeley, USA, 2020, pp. 486-489.

[16] P. Xu, Q. Gu. A finite-time analysis of q-learning with neural network function approxima-tion. Proceedings of the 37th International Conference on Machine Learning. Virtual, 2020, pp. 10555-10565.

[17] V. Mnih, P. A Badia, M. Mirza, et al. Asynchronous methods for deep reinforcement learning. Proceedings of the 33rd International Conference on Machine Learning. NY, USA, 2016, pp. 1928-1937.

[18] W. Fedus, P. Ramachandran, R. Agarwal, et al. Revisiting fundamentals of experience replay. Proceedings of the 37th International Conference on Machine Learning. Virtual, 2020, pp. 3061-3071.

[19] J. S Reddi, S. Kale, S. Kumar. On the convergence of adam and beyond. Proceedings of the 6th International Conference on Learning Representations. BC, Canada, 2018.

[20] Q. Cai, Z. Yang, D. J Lee, et al. Neural temporal-difference learning converges to global opti-ma. Proceedings of the 33rd International Conference on Neural Information Processing Systems. BC, Canada, 2019, pp. 11312-11322.