

Improved and Stable Data-Based Tracking Control via Approximate Iterative Q-Learning

Zhantao Liang

School of Automation
Guangdong University of Technology
Guangzhou, China
2112104135@mail2.gdut.edu.cn

Weiming Liu

School of System Design and Intelligent Manufacturing
Southern University of Science and Technology
Shenzhen, China
weiming.leo@foxmail.com

Mingming Ha

School of Intelligence Science and Technology
and Institute of Artificial Intelligence
University of Science and Technology Beijing
Beijing, China
hamingming_0705@foxmail.com

Derong Liu*

School of System Design and Intelligent Manufacturing
Southern University of Science and Technology
Shenzhen, China
liudr@sustech.edu.cn

Abstract—In this paper, to avoid computing the feedward control input and to completely eliminate the tracking error, an improved cost function with respect to the tracking error is proposed, under which the optimal tracking controller can stabilize the tracking error system. Then, the approximate value-iteration-based Q-learning is given to solve the neuro-optimal control policy. Also, the finite error bound of the approximate Q-function is established. In addition, for guaranteeing the admissibility of the obtained control policies, some stability conditions for this data-based iterative Q-learning are developed. When the stability conditions are satisfied, it is ensured that the iterative control policies derived from the critic network and the trained action network with approximation errors will stabilize the tracking error system. Finally, a simulation example is employed to implement the data-based Q-learning algorithm and verify the present theoretical results.

Index Terms—Adaptive dynamic programming, discrete-time nonlinear systems, value-iteration-based Q-learning, data-based tracking control, stability analysis

I. INTRODUCTION

Being characterized by self-adaptivity and self-learning, adaptive dynamic programming (ADP) has been concerned and developed in the domain of optimal control in recently years [1], [2]. To overcome the “curse of dimensionality” caused by DP, ADP adopts the neural networks (NNs) to approximate the cost function and control policy, called critic network and action network. Based on the Q-function approximated by the critic network, one of the iterative ADP algorithm is Q-learning, which do not need the knowledge of the system model and can get the optimal control policy through off-policy learning. However, in iterative Q-learning, the sequence of the iterative Q-function converges to the optimal Q-function provided that it can be computed exactly. When NN is adopted to approximate the iterative Q-function, in fact, the resulting approximation errors will influence the

convergence of the sequence of that and the admissibility of the obtained control policy. In [3], the finite error bound of the approximate Q-function derived from policy iteration (PI) was established to guarantee the sequence ultimately ‘converges’ to the neighborhood of the optimal Q-function under a mild boundedness condition. Moreover, in [4], under the approximate value-iteration-based Q-learning (AVIQL), some stability conditions were developed to guarantee the iterative control policy derived from the obtained critic network or the trained action network with approximation errors makes the controlled system asymptotically stable.

As a class of optimal control problems, the optimal tracking control is to design the optimal tracking controller that makes the system state track the reference trajectory with the minimum cost. In general method, the tracking controller is divided into feedforward control and feedback control, which are designed, respectively [5], [6]. However, the computation of feedforward control generally requires the knowledge of system model and is not unique. Another method is to solve the tracking controller by minimizing the cost function for the augmented system [6]. However, since the augmented system is not controllable, the tracking error may not be eliminated completely under the optimal control policy [7], [8]. In the view of these shortcomings, a novel cost function was proposed to eliminate the tracking error without computing the feedforward control, where the control input is introduced into the tracking error and the quadratic form of it is ignored [9]. Moreover, the new stability analysis was presented for guaranteeing the tracking error is eliminated completely under the discounted VI-based critic framework, which is different from the Lyapunov method in the optimal state regulator problem [10].

In this paper, for obtaining the tracking controller that stabilizes the tracking error system, an improved cost function with respect to the tracking error and its difference is proposed.

* Corresponding author.

The AVIQL is applied to solve the neuro-optimal control policy and the finite error bound of the approximate Q-function is given. Additionally, considering the approximation errors caused by the critic and action network, some stability conditions are developed to guarantee the stability of the tracking error system under the iterative control policy and the trained action network.

II. PRELIMINARY

We consider the discrete-time nonlinear system described as

$$x_{k+1} = \mathcal{F}(x_k, u_k), \quad (1)$$

where $x_k \in \mathbb{R}^n$ and $u_k \in \mathbb{R}^m$ are the system state and control input at the time step $k \in \mathbb{N}$, and $\mathcal{F}: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ is the continuous system function.

We also consider the reference trajectory generated by

$$r_{k+1} = \mathcal{G}(r_k), \quad (2)$$

where $r_k \in \mathbb{R}^n$ is the reference signal at k , and $\mathcal{G}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the continuous command generator. Defining the tracking error as

$$e_k = x_k - r_k, \quad (3)$$

and the tracking error system as

$$e_{k+1} = \mathcal{F}(e_k + r_k, u_k) - \mathcal{G}(r_k), \quad (4)$$

we assume there exists an infinite-length control input sequence $\underline{e}_0 = \{u_0, u_1, \dots\}$ such that $e_k \rightarrow 0$ as $k \rightarrow \infty$ for the tracking error system (4). The goal of the tracking control problem is to design a tracking controller $u_k = \mu(e_k, r_k)$ that stabilizes (4), where $\mu: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the continuous control policy. To avoid computing feedforward control input and to eliminate the tracking error completely, we develop an improved cost function with the discount factor $\gamma \in (0, 1]$ as

$$\begin{aligned} \mathcal{J}(e_0, r_0, \underline{e}_0) &= \sum_{k=0}^{\infty} \gamma^k U(e_k, r_k, u_k) \\ &= \sum_{k=0}^{\infty} \gamma^k (Q(e_k) + R(e_{k+1} - e_k)), \end{aligned} \quad (5)$$

where $U: \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ is the continuous utility function, $Q: \mathbb{R}^n \rightarrow \mathbb{R}_+$ and $R: \mathbb{R}^n \rightarrow \mathbb{R}_+$ are both positive definite function. We assume there exists an admissible control policy for the tracking error system (4) and the discounted cost function (5). Note that the effect of γ in (5) is different from that in the cost function for augmented systems. Besides, if γ satisfies some conditions [11, Theorem 2], the optimal control policy of (5) can eliminates the the tracking error completely.

Under an admissible control policy $\mu(e_k, r_k)$, the value function is defined as

$$\begin{aligned} \mathcal{V}^\mu(e_k, r_k) &= \mathcal{J}(e_k, r_k, \mu(e_k, r_k)) \\ &= \sum_{l=0}^{\infty} \gamma^{k+l} U(e_{k+l}, r_{k+l}, \mu(e_{k+l}, r_{k+l})). \end{aligned} \quad (6)$$

The optimal value function is obtained by

$$\mathcal{V}^*(e_k, r_k) = \inf_{\mu} \mathcal{V}^\mu(e_k, r_k). \quad (7)$$

According to the principle of optimality, $\mathcal{V}^*(e_k, r_k)$ satisfies the following Bellman equation

$$\mathcal{V}^*(e_k, r_k) = \min_{u_k} \{U(e_k, r_k, u_k) + \gamma \mathcal{V}^*(e_{k+1}, r_{k+1})\}. \quad (8)$$

The optimal control policy is obtained by

$$\mu^*(e_k, r_k) = \arg \min_{u_k} \{U(e_k, r_k, u_k) + \gamma \mathcal{V}^*(e_{k+1}, r_{k+1})\}. \quad (9)$$

Furthermore, the Q-function for $\mu(e_k, r_k)$ is defined as

$$\begin{aligned} \mathcal{Q}^\mu(e_k, r_k, u_k) &= U(e_k, r_k, u_k) + \gamma \mathcal{V}^\mu(e_{k+1}, r_{k+1}) \\ &= U(e_k, r_k, u_k) \\ &\quad + \gamma \mathcal{Q}^\mu(e_{k+1}, r_{k+1}, \mu(e_{k+1}, r_{k+1})). \end{aligned} \quad (10)$$

The optimal Q-function is obtained by

$$\mathcal{Q}^*(e_k, r_k, u_k) = \min_{\mu} \mathcal{Q}^\mu(e_k, r_k, u_k). \quad (11)$$

According to [12] and [13], $\mathcal{Q}^*(e_k, r_k, u_k)$ satisfies the optimality equation

$$\mathcal{Q}^*(e_k, r_k, u_k) = U(e_k, r_k, u_k) + \gamma \min_{u_{k+1}} \mathcal{Q}^*(e_{k+1}, r_{k+1}, u_{k+1}). \quad (12)$$

The optimal control policy is also obtained by

$$\mu^*(e_k, r_k) = \arg \min_{u_k} \mathcal{Q}^*(e_k, r_k, u_k). \quad (13)$$

III. DATA-BASED AVIQL FOR TRACKING CONTROL

As a form of model-free reinforcement learning (RL), Q-learning allows the agent to learn the optimal control policy by off-policy. In this paper, we introduce the (approximate) VIQL to solve the (near) optimal tracking controller of (5).

A. VIQL Algorithm

In VIQL, starting with a positive definite initial Q-function $\mathcal{Q}_0(e, r, u)$, the initial control policy is obtained by

$$\mu_0(e', r') = \arg \min_{u'} \mathcal{Q}_0(e', r', u'), \quad (14)$$

where

$$e' = \mathcal{F}(e + r, u) - \mathcal{G}(r),$$

$r' = \mathcal{G}(r)$ and u' is the control input at the next time step. Then for $i \in \mathbb{N}_+$, VIQL iterates between value function update

$$\begin{aligned} \mathcal{Q}_i(e, r, u) &= U(e, r, u) + \gamma \min_{u'} \mathcal{Q}_{i-1}(e', r', u') \\ &= U(e, r, u) + \gamma \mathcal{Q}_{i-1}(e', r', \mu_{i-1}(e', r')), \end{aligned} \quad (15)$$

and control policy improvement

$$\mu_i(e', r') = \arg \min_{u'} \mathcal{Q}_i(e', r', u'), \quad (16)$$

where $\mathcal{Q}_i(e, r, u)$ and $\mu_i(e', r')$ are the iterative Q-function and control policy.

Theorem 1: Assume $0 \leq \mathcal{Q}^*(e, r, u) \leq (1 + \eta)U(e, r, u)$ holds uniformly for some $\eta \in (0, \infty)$ and $0 \leq \alpha \mathcal{Q}^*(e, r, u) \leq \mathcal{Q}_0(e, r, u) \leq \beta \mathcal{Q}^*(e, r, u)$, $\alpha \in [0, 1]$ and $\beta \in [1, \infty)$. And the

VIQL is given by (15) and (16). Then, for $i \in \mathbb{N}$, $\mathcal{Q}_i(e, r, u)$ satisfies

$$\begin{aligned} \left[1 + \frac{\alpha - 1}{(1 + 1/\eta)^i}\right] \mathcal{Q}^*(e, r, u) &\leq \mathcal{Q}_i(e, r, u) \\ &\leq \left[1 + \frac{\beta - 1}{(1 + 1/\eta)^i}\right] \mathcal{Q}^*(e, r, u). \end{aligned} \quad (17)$$

Moreover, as $i \rightarrow \infty$, we have

$$\lim_{i \rightarrow \infty} \mathcal{Q}_i(e, r, u) = \mathcal{Q}^*(e, r, u). \quad (18)$$

Proof: According to (12) and the assumption that $0 \leq \mathcal{Q}^*(e, r, u) \leq (1 + \eta)U(e, r, u)$ for some $\eta \in (0, \infty)$, it gives

$$\gamma \min_{u'} \mathcal{Q}^*(e', r', u') \leq \eta U(e, r, u). \quad (19)$$

(17) can be shown by mathematical induction. For $i = 0$, obviously $\mathcal{Q}_0(e, r, u)$ satisfies the left-hand side of (17) since $\mathcal{Q}_0(e, r, u) \geq \alpha \mathcal{Q}^*(e, r, u)$, $\alpha \in [0, 1]$. Assume for $i = l$, $\mathcal{Q}_l(e, r, u)$ satisfies the left-hand side of (17). Then for $i = l + 1$, it leads to

$$\begin{aligned} \mathcal{Q}_{l+1}(e, r, u) &= U(e, r, u) + \gamma \min_{u'} \mathcal{Q}_l(e', r', u') \\ &\geq \left[1 + \frac{\alpha - 1}{(1 + 1/\eta)^{l+1}}\right] U(e, r, u) \\ &\quad + \left[1 + \frac{\alpha - 1}{(1 + 1/\eta)^l} - \frac{(\alpha - 1)/\eta}{(1 + 1/\eta)^{l+1}}\right] \\ &\quad \times \gamma \min_{u'} \mathcal{Q}^*(e', r', u') \\ &= \left[1 + \frac{\alpha - 1}{(1 + 1/\eta)^{l+1}}\right] \mathcal{Q}^*(e, r, u), \end{aligned} \quad (20)$$

which satisfies the left-hand side of (17). The mathematical induction is complete. Similar method is used to get the right-hand side of (17). Taking limiting of the both-hand side of (17), one has (18). ■

B. Derivation of AVIQL

Numerically computing the iterative Q-function through discretization method always leads to ‘curse of dimensionality’. In ADP, some off-policy data are used to compute their values and the NN is adopted to approximate the Q-function. However, for general nonlinear systems, the approximation errors caused by the critic network cannot be ignored.

AVIQL starts from the initial Q-function $\hat{\mathcal{Q}}_0(e, r, u) = \mathcal{Q}_0(e, r, u)$ and computes the initial control policy by

$$\tilde{\mu}_0(e', r') = \arg \min_{u'} \hat{\mathcal{Q}}_0(e', r', u'). \quad (21)$$

Then it is iteratively implemented between

$$\begin{aligned} \hat{\mathcal{Q}}_i(e, r, u) &= U(e, r, u) + \gamma \min_{u'} \hat{\mathcal{Q}}_{i-1}(e', r', u') \\ &\quad + \rho_{i-1}(e, r, u) \\ &= U(e, r, u) + \gamma \hat{\mathcal{Q}}_{i-1}(e', r', \tilde{\mu}_{i-1}(e', r')) \\ &\quad + \rho_{i-1}(e, r, u) \end{aligned} \quad (22)$$

and

$$\tilde{\mu}_i(e', r') = \arg \min_{u'} \hat{\mathcal{Q}}_i(e', r', u'), \quad (23)$$

$\forall i \in \mathbb{N}_+$, where $\hat{\mathcal{Q}}_i(e, r, u)$ is the approximate Q-function evaluating $\tilde{\mu}_i(e, r)$, $\tilde{\mu}_i(e', r')$ is the improved control policy derived from $\hat{\mathcal{Q}}_i(e', r', u')$, and $\rho_{i-1}(e, r, u)$ is the approximation errors of the Q-function. Note that the iterative control policy is derived from the critic network. However, when NN is adopted to approximate it, the approximation errors caused by the action network cannot be ignored.

Theorem 2: Assume $|\rho_i(e, r, u)| \leq cU(e, r, u)$, $\forall i \in \mathbb{N}$, holds for some $c \in [0, 1)$. And the AVIQL is given by (22) and (23). Then, for $i \in \mathbb{N}$, $\hat{\mathcal{Q}}_i(e, r, u)$ satisfies

$$\begin{aligned} (1 - c) \left[1 + \frac{\alpha - 1}{(1 + 1/\eta)^i}\right] \mathcal{Q}^*(e, r, u) &\leq \hat{\mathcal{Q}}_i(e, r, u) \\ &\leq (1 + c) \left[1 + \frac{\beta - 1}{(1 + 1/\eta)^i}\right] \mathcal{Q}^*(e, r, u). \end{aligned} \quad (24)$$

Moreover, as $i \rightarrow \infty$, we have

$$(1 - c) \mathcal{Q}^*(e, r, u) \leq \lim_{i \rightarrow \infty} \hat{\mathcal{Q}}_i(e, r, u) \leq (1 + c) \mathcal{Q}^*(e, r, u). \quad (25)$$

Proof: (24) can be shown by mathematical induction. For $i = 0$, obviously $\hat{\mathcal{Q}}_0(e, r, u)$ satisfies the right-hand side of (24) since $\mathcal{Q}_0(e, r, u) \leq \beta \mathcal{Q}^*(e, r, u)$, $\beta \in [1, \infty)$. Assume for $i = l$, $\hat{\mathcal{Q}}_l(e, r, u)$ satisfies the right-hand side of (24). Then for $i = l + 1$, due to $|\rho_i(e, r, u)| \leq cU(e, r, u)$ for some $c \in [0, 1)$, combining (19), it leads to

$$\begin{aligned} \hat{\mathcal{Q}}_{l+1}(e, r, u) &= U(e, r, u) + \gamma \min_{u'} \hat{\mathcal{Q}}_l(e', r', u') + \rho_l(e, r, u) \\ &\leq (1 + c) \left[\left(1 + \frac{\beta - 1}{(1 + 1/\eta)^{l+1}}\right) U(e, r, u) \right. \\ &\quad \left. + \left(1 + \frac{\beta - 1}{(1 + 1/\eta)^l} - \frac{(\beta - 1)/\eta}{(1 + 1/\eta)^{l+1}}\right) \right. \\ &\quad \left. \times \gamma \min_{u'} \mathcal{Q}^*(e', r', u') \right] \\ &= (1 + c) \left[1 + \frac{\beta - 1}{(1 + 1/\eta)^{l+1}}\right] \mathcal{Q}^*(e, r, u), \end{aligned} \quad (26)$$

which satisfies the right-hand side of (24). The mathematical induction is complete. Similar method is used to get the right-hand side of (24). Taking limiting of the both-hand side of (24), one has (25). ■

IV. NOVEL STABILITY ANALYSIS OF AVIQL FOR TRACKING CONTROL

A. Admissibility of the Iterative Control Policy

Since the optimal control policy is not always achieved for general nonlinear systems, the iterative control policy is usually applied to replace it. However, considering the approximation errors of the critic network, the admissibility of the iterative control policy derived from the critic network should be reconsidered. Define the iterative function as

$$\delta_i(e, r, u) = |\hat{\mathcal{Q}}_{i+1}(e, r, u) - \gamma \hat{\mathcal{Q}}_i(e, r, u)|. \quad (27)$$

When $\delta_i(e, r, u)$ satisfies the stability condition, $\tilde{\mu}_i(e, r, u) = \arg \min_{u'} \hat{\mathcal{Q}}_i(e, r, u)$ is applied to control the system with stability guarantee.

Theorem 3: Assume $|\rho_i(e, r, u)| \leq cU(e, r, u)$, $\forall i \in \mathbb{N}$, holds for some $c \in [0, 1)$. And the AVIQL is given by (22) and (23). If $\delta_i(e, r, u)$ satisfies

$$\delta_i(e, r, u) \leq (1 - a)U(e, r, u), \quad (28)$$

where $a \in (c, 1)$ is a constant, then $\tilde{\mu}_i(e, r)$ is an admissible control policy.

Proof: According to (22), (27) and (28), one has

$$aU(e, r, u) + \rho_i(e, r, u) \leq \gamma \hat{Q}_i(e, r, u) - \gamma \hat{Q}_i(e', r', \tilde{\mu}_i(e', r')). \quad (29)$$

Due to $|\rho_i(e, r, u)| \leq cU(e, r, u)$, (29) becomes

$$\frac{a-c}{\gamma}U(e, r, u) \leq \hat{Q}_i(e, r, u) - \hat{Q}_i(e', r', \tilde{\mu}_i(e', r')). \quad (30)$$

Substituting $u = \tilde{\mu}_i(e, r)$ into (30), one has

$$\begin{aligned} \frac{a-c}{\gamma}U(e, r, \tilde{\mu}_i(e, r)) &\leq \hat{Q}_i(e, r, \tilde{\mu}_i(e, r)) \\ &\quad - \hat{Q}_i(e', r', \tilde{\mu}_i(e', r')), \end{aligned} \quad (31)$$

where $\tilde{e}' = \mathcal{F}(e + r, \tilde{\mu}_i(e, r)) - \mathcal{G}(r)$. Let $\tilde{e}_{k+1} = \mathcal{F}(\tilde{e}_k + r_k, \tilde{\mu}_i(\tilde{e}_k, r_k)) - \mathcal{G}(r)$, $\forall k \in \mathbb{N}$, and $\tilde{e}_0 = e_0$. Evaluating (31) at $\tilde{e}_k$, $\forall k \in \mathbb{N}$, it leads to

$$\begin{aligned} \frac{a-c}{\gamma}U(\tilde{e}_k, r_k, \tilde{\mu}_i(\tilde{e}_k, r_k)) &\leq \hat{Q}_i(\tilde{e}_k, r_k, \tilde{\mu}_i(\tilde{e}_k, r_k)) \\ &\quad - \hat{Q}_i(\tilde{e}_{k+1}, r_{k+1}, \tilde{\mu}_i(\tilde{e}_{k+1}, r_{k+1})). \end{aligned} \quad (32)$$

By accumulating (32) for K times in sequence, it results in

$$\begin{aligned} 0 &\leq \frac{a-c}{\gamma} \sum_{k=0}^K U(\tilde{e}_k, r_k, \tilde{\mu}_i(\tilde{e}_k, r_k)) \\ &\leq \hat{Q}_i(\tilde{e}_0, r_0, \tilde{\mu}_i(\tilde{e}_0, r_0)) \\ &\quad - \hat{Q}_i(\tilde{e}_{K+1}, r_{K+1}, \tilde{\mu}_i(\tilde{e}_{K+1}, r_{K+1})) \\ &\leq \hat{Q}_i(\tilde{e}_0, r_0, \tilde{\mu}_i(\tilde{e}_0, r_0)). \end{aligned} \quad (33)$$

Letting $K \rightarrow \infty$ in (33), one has

$$0 \leq \sum_{k=0}^{\infty} U(\tilde{e}_k, r_k, \tilde{\mu}_i(\tilde{e}_k, r_k)) \leq \frac{\gamma}{a-c} \hat{Q}_i(\tilde{e}_0, r_0, \tilde{\mu}_i(\tilde{e}_0, r_0)). \quad (34)$$

Since $\hat{Q}_i(\tilde{e}_0, r_0, \tilde{\mu}_i(\tilde{e}_0, r_0))$ is a finite function, it implies that $\sum_{k=0}^{\infty} U(\tilde{e}_k, r_k, \tilde{\mu}_i(\tilde{e}_k, r_k)) < \infty$. In addition, since $U(\tilde{e}_k, r_k, \tilde{\mu}_i(\tilde{e}_k, r_k)) = Q(\tilde{e}_k) + R(\tilde{e}_{k+1} - \tilde{e}_k)$, where $Q(\cdot)$ and $R(\cdot)$ are positive definite functions, it means that $Q(\tilde{e}_k) \rightarrow 0$ as $k \rightarrow \infty$. Hence, one has $\tilde{e}_k \rightarrow 0$ as $k \rightarrow \infty$. ■

B. Admissibility of the Action Network

To reduce the computing burden and increase the real-time performance, we adopt another NN to approximate the iterative control policy derived from the critic network which satisfies (28). Considering the approximation errors caused by the action network, the admissibility of the action network should be reanalyzed. Denote the approximation of $\tilde{\mu}_i(e, r)$ as

$$\hat{\mu}_i(e, r) = \tilde{\mu}_i(e, r) + \sigma_i(e, r), \quad (35)$$

where $\sigma_i(e, r)$ is the approximation errors of the control policy. When $\sigma_i(e, r)$ is lower than the upper bound function, $\hat{\mu}_i(e, r)$ is applied to control the system with stability guarantee.

Theorem 4: Assume $|\rho_i(e, r, u)| \leq cU(e, r, u)$, $\forall i \in \mathbb{N}$, holds for some $c \in [0, 1)$. And the AVIQL is given by (22) and (23). Also the Lipschitz constant of function $\hat{Q}_i(e, r, \cdot)$ is given by L_Q . If $\sigma_i(e, r)$ satisfies

$$\|\sigma_i(e, r)\| \leq \min_u \frac{(1-a)U(e, r, u) - \delta_i(e, r, u)}{\gamma L_Q}, \quad (36)$$

where $a \in (c, 1)$ is a constant, then $\hat{\mu}_i(e, r)$ is an admissible control policy.

Proof: Considering the smoothness of the function approximator, $\hat{Q}_i(e, r, \cdot)$ is Lipschitz continuous on a compact subset $\mathcal{U} \subset \mathbb{R}^m$, which means

$$\|\hat{Q}_i(e, r, u) - \hat{Q}_i(e, r, v)\| \leq L_Q \|u - v\|, \quad (37)$$

$\forall u, v \in \mathcal{U}$. Combining (35), (37) becomes

$$\hat{Q}_i(e, r, \hat{\mu}_i(e, r)) - L_Q \|\sigma_i(e, r)\| \leq \hat{Q}_i(e, r, \tilde{\mu}_i(e, r)). \quad (38)$$

According to (27) and (36), one has

$$\begin{aligned} aU(e, r, u) + \rho_i(e, r, u) &\leq \gamma \hat{Q}_i(e, r, u) \\ &\quad - \gamma \hat{Q}_i(e', r', \tilde{\mu}_i(e', r')) \\ &\quad - \gamma L_Q \|\sigma_i(e, r)\|. \end{aligned} \quad (39)$$

Due to $|\rho_i(e, r, u)| \leq cU(e, r, u)$, (39) becomes

$$\begin{aligned} \frac{a-c}{\gamma}U(e, r, u) &\leq \hat{Q}_i(e, r, u) - \hat{Q}_i(e', r', \tilde{\mu}_i(e', r')) \\ &\quad - L_Q \|\sigma_i(e, r)\|. \end{aligned} \quad (40)$$

Substituting $u = \hat{\mu}_i(e, r)$ into (40), one has

$$\begin{aligned} \frac{a-c}{\gamma}U(e, r, \hat{\mu}_i(e, r)) &\leq \hat{Q}_i(e, r, \hat{\mu}_i(e, r)) \\ &\quad - \hat{Q}_i(e', r', \tilde{\mu}_i(e', r')) \\ &\quad - L_Q \|\sigma_i(e, r)\|. \end{aligned} \quad (41)$$

where $\tilde{e}' = \mathcal{F}(e + r, \hat{\mu}_i(e, r)) - \mathcal{G}(r)$. Using (38) to replace $\hat{Q}_i(e, r, \hat{\mu}_i(e, r))$ in (41) with $\hat{Q}_i(e, r, \tilde{\mu}_i(e, r))$, it leads to

$$\begin{aligned} \frac{a-c}{\gamma}U(e, r, \hat{\mu}_i(e, r)) &\leq \hat{Q}_i(e, r, \tilde{\mu}_i(e, r)) \\ &\quad - \hat{Q}_i(e', r', \tilde{\mu}_i(e', r')). \end{aligned} \quad (42)$$

Let $\hat{e}_{k+1} = \mathcal{F}(\hat{e}_k + r_k, \hat{\mu}_i(\hat{e}_k, r_k)) - \mathcal{G}(r_k)$, $\forall k \in \mathbb{N}$, and $\hat{e}_0 = e_0$. Evaluating (42) at $\hat{e}_k$, $\forall k \in \mathbb{N}$, it leads to

$$\begin{aligned} \frac{a-c}{\gamma}U(\hat{e}_k, r_k, \hat{\mu}_i(\hat{e}_k, r_k)) &\leq \hat{Q}_i(\hat{e}_k, r_k, \tilde{\mu}_i(\hat{e}_k, r_k)) \\ &\quad - \hat{Q}_i(\hat{e}_{k+1}, r_{k+1}, \tilde{\mu}_i(\hat{e}_{k+1}, r_{k+1})). \end{aligned} \quad (43)$$

By accumulating (44) for K times in sequence, it results in

$$\begin{aligned} 0 &\leq \frac{a-c}{\gamma} \sum_{k=0}^K U(\hat{e}_k, r_k, \hat{\mu}_i(\hat{e}_k, r_k)) \\ &\leq \hat{Q}_i(\hat{e}_0, r_0, \tilde{\mu}_i(\hat{e}_0, r_0)) \\ &\quad - \hat{Q}_i(\hat{e}_{K+1}, r_{K+1}, \tilde{\mu}_i(\hat{e}_{K+1}, r_{K+1})) \\ &\leq \hat{Q}_i(\hat{e}_0, r_0, \tilde{\mu}_i(\hat{e}_0, r_0)). \end{aligned} \quad (44)$$

Letting $K \rightarrow \infty$ in (44), one has

$$0 \leq \sum_{k=0}^{\infty} U(\hat{e}_k, r_k, \hat{\mu}_i(\hat{e}_k, r_k)) \leq \frac{\gamma}{a-c} \hat{Q}_i(\hat{e}_0, r_0, \hat{\mu}_i(\hat{e}_0, r_0)). \quad (45)$$

Since $\hat{Q}_i(\hat{e}_0, r_0, \hat{\mu}_i(\hat{e}_0, r_0))$ is a finite function, it implies that $\sum_{k=0}^{\infty} U(\hat{e}_k, r_k, \hat{\mu}_i(\hat{e}_k, r_k)) < \infty$. In addition, since $U(\hat{e}_k, r_k, \hat{\mu}_i(\hat{e}_k, r_k)) = Q(\hat{e}_k) + R(\hat{e}_{k+1} - \hat{e}_k)$, where $Q(\cdot)$ and $R(\cdot)$ are positive definite functions, it means that $Q(\hat{e}_k) \rightarrow 0$ as $k \rightarrow \infty$. Hence, one has $\hat{e}_k \rightarrow 0$ as $k \rightarrow \infty$. ■

V. SIMULATION STUDIES

We consider the modified Van der Pol's oscillator system [14] described as

$$\dot{x} = \begin{bmatrix} x_2 \\ (1 - x_1^2)x_2 - x_1 \end{bmatrix} + Bu, \quad (46)$$

where $B = 3.5I_2$. Let the sampling interval be $\Delta t = 0.05$ sec. The discrete-time state equation is given by

$$x_{k+1} = \begin{bmatrix} x_{1k} + 0.05x_{2k} \\ x_{2k} + 0.05((1 - x_{1k}^2)x_{2k} - x_{1k}) \end{bmatrix} + 0.05Bu_k. \quad (47)$$

Assume for (47), lots of data is collected from daily operation and recorded as the form $\langle x, u, x' \rangle$, where $x' = \mathcal{F}(x, u)$. Define the sample set as $\mathcal{X} = \{\langle x_{[d]}, u_{[d]}, x'_{[d]} \rangle | d \in \mathbb{N}_+\}$. We consider the reference trajectory is generated by

$$r_{k+1} = \begin{bmatrix} \cos(0.05) & \sin(0.05) \\ -\sin(0.05) & \cos(0.05) \end{bmatrix} r_k. \quad (48)$$

Assume for (48), some samples of the desired trajectory are recorded as the form $\langle r, r' \rangle$. Let $\mathcal{R} = \{\langle r_{[d]}, r'_{[d]} \rangle | d \in \mathbb{N}_+\}$. Select 2000 data from $\mathcal{X}$ and $\mathcal{R}$ to construct the data set as $\mathcal{T} = \{\langle e_{[d]}, r_{[d]}, u_{[d]}, e'_{[d]}, r'_{[d]} \rangle | \langle x_{[d]}, u_{[d]}, x'_{[d]} \rangle \in \mathcal{X}, \langle r_{[d]}, r'_{[d]} \rangle \in \mathcal{R}\}$. One part of $\mathcal{T}$ serves as the training set $\mathcal{T}_{ra}$, which has 1500 samples. The other one part of $\mathcal{T}$ serves as the test set $\mathcal{T}_{es}$, which has 500 samples.

Select the utility function as

$$U(e, r, u) = e^T Q e + (e' - e)^T R (e' - e),$$

where $Q = I_2$ and $R = 20I_2$. Select the critic network as three-layer NN with structure 6-27-1, which is expressed as

$$\hat{Q}_i(e, r, u) = w_{c1,i}^T \phi_c(w_{c2,i}^T [e^T, r^T, u^T]^T + b_{c1,i}) + b_{c2,i}, \quad (49)$$

where $\phi_c(\cdot) = \tanh(\cdot)$ is the activation function, $w_{c1,i}$, $w_{c2,i}$, $b_{c1,i}$ and $b_{c2,i}$ are the parameters with appropriate dimensions.

Let $\gamma = 1$ and $\hat{Q}_0(e, r, u) = 0$. In each iteration, $\mathcal{T}_{ra}$ is used to learn the critic network through (22) and (23). The AVIQL is terminated at 16-th iteration with $|\hat{Q}_{18}(e, r, u) - \gamma \hat{Q}_{17}(e, r, u)| \leq 0.001$. To evaluate the approximation accuracy and verify the admissibility of $\hat{\mu}_{17}(e, r)$, $\mathcal{T}_{es}$ is used to test. The approximation errors $\rho_i(e_{[d]}, r_{[d]}, u_{[d]})$,

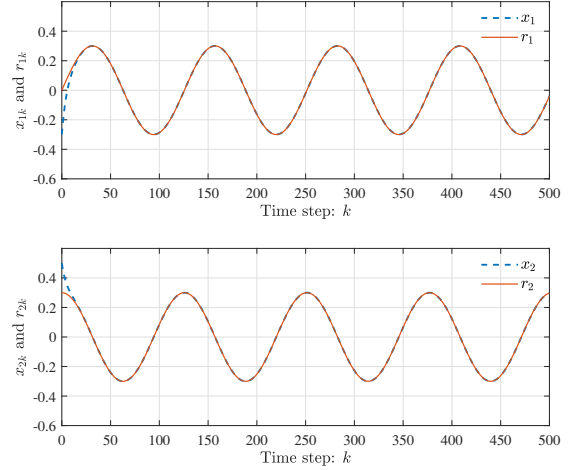


Fig. 1. The system state and reference trajectories under $\hat{\mu}_{17}(e, r)$.

$\forall \langle e_{[d]}, r_{[d]}, u_{[d]}, e'_{[d]}, r'_{[d]} \rangle \in \mathcal{T}_{es}, \forall i$ is given by (22). The error coefficient c can be estimated by

$$c \approx \max_{\forall \langle e_{[d]}, r_{[d]}, u_{[d]}, e'_{[d]}, r'_{[d]} \rangle \in \mathcal{T}_{es}, \forall i} \frac{\rho_i(e_{[d]}, r_{[d]}, u_{[d]})}{U(e_{[d]}, r_{[d]}, u_{[d]})} = 0.0605.$$

The function $\delta_{17}(e_{[d]}, r_{[d]}, u_{[d]})$, $\forall \langle e_{[d]}, r_{[d]}, u_{[d]} \rangle \in \mathcal{T}_{es}$, is given by (27). Through pointwise comparison, we have $\delta_{17}(e_{[d]}, r_{[d]}, u_{[d]}) \leq (1-a)U(e_{[d]}, r_{[d]}, u_{[d]})$, $\forall \langle e_{[d]}, r_{[d]}, u_{[d]} \rangle \in \mathcal{T}_{es}$. The stability of the tracking error system under $\hat{\mu}_{17}(e, r)$ follows. Select $x_0 = [-0.3, 0.5]^T$ and $r_0 = [0, 0.3]^T$ and apply $\hat{\mu}_{17}(e, r)$ to control the system. The system state and reference signal are plotted in Fig. 3, and the control input and tracking error are plotted in Fig. 4, versus the time steps.

Consider the action network being adopted to approximate $\hat{\mu}_{17}(e, r)$, which is selected as three-layer NN with structure 4-21-2 and formulated as

$$\hat{\mu}_{17}(e, r) = w_{a1}^T \phi_a(w_{a2}^T [e^T, r^T]^T + b_{a1}) + b_{a2}, \quad (50)$$

where $\phi_a(\cdot) = \tanh(\cdot)$ is the activation function, w_{a1} , w_{a2} , b_{a1} and b_{a2} are the parameters with appropriate dimensions. $\mathcal{T}_{ra}$ is used to trained the action network by (35) and $\mathcal{T}_{es}$ is used to verify the admissibility of $\hat{\mu}_{17}(e, r)$ by (36). The approximation errors $\sigma_{17}(e_{[d]}, r_{[d]})$, $\forall \langle e_{[d]}, r_{[d]} \rangle \in \mathcal{T}_{es}$, is given by (35). The Lipschitz constant L_Q is estimated by

$$L_Q \approx \max_{\forall \langle e_{[d]}, r_{[d]}, u_{[d]} \rangle \in \mathcal{T}_{es}} \left\| \frac{\partial \hat{Q}_{17}(e_{[d]}, r_{[d]}, u_{[d]})}{\partial u} \right\| = 2.6513.$$

Through pointwise comparison, we have $\|\sigma_{17}(e_{[d]}, r_{[d]})\| \leq [(1-a)U(e_{[d]}, r_{[d]}, u_{[d]}) - \delta_{17}(e_{[d]}, r_{[d]}, u_{[d]})] / (\gamma L_Q)$, $\forall \langle e_{[d]}, r_{[d]}, u_{[d]}, e'_{[d]} \rangle \in \mathcal{T}_{es}$. The stability of the tracking error system under $\hat{\mu}_{17}(e, r)$ follows. Select $x_0 = [0.3, 0]^T$ and $r_0 = [0, 0.3]^T$, and apply $\hat{\mu}_{17}(e, r)$ to control the system. The system state and reference signal are plotted in Fig. 3, and the control input and tracking error are plotted in Fig. 4, versus the time steps.

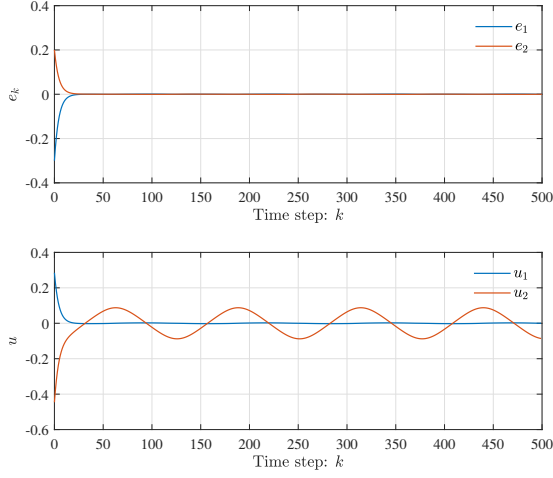


Fig. 2. The control input and tracking error trajectories under $\tilde{\mu}_{17}(e, r)$.

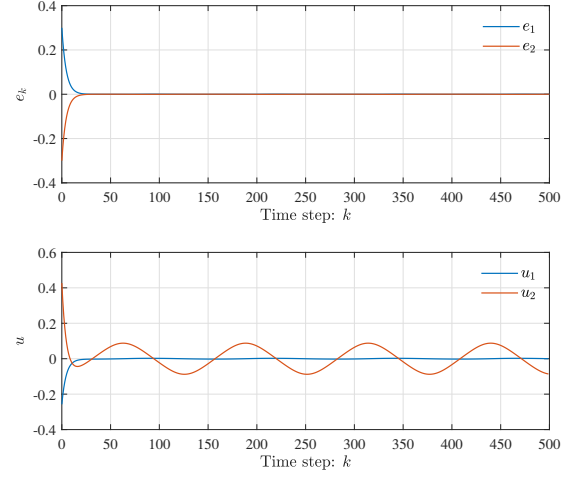


Fig. 4. The control input and tracking error trajectories under $\hat{\mu}_{17}(e, r)$.

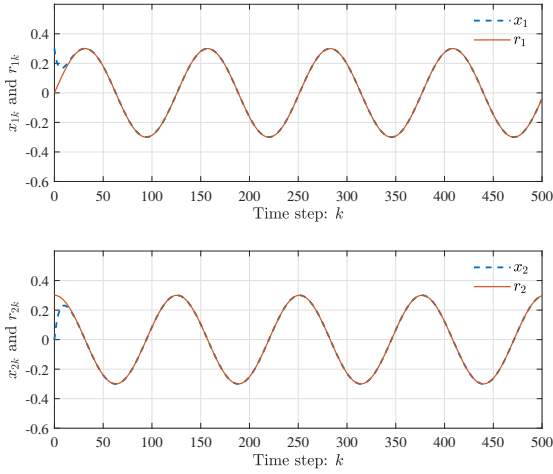


Fig. 3. The system state and reference trajectories under $\hat{\mu}_{17}(e, r)$.

VI. CONCLUSIONS

In this paper, for the data-based tracking control, an improved cost function with respect to the tracking error is proposed, under which the optimal tracking controller stabilizes the tracking error system. Then, the AVIQL is applied to solve the neuro-optimal tracking controller using data and NNs. Also, the finite error bound of the approximate Q-function is given based on the mild boundedness condition. Furthermore, for guaranteeing the admissibility of the obtained control policies, the stability condition about the critic network and the upper bound function of the approximation errors of the action network are developed. It is ensured that the iterative control policy and the trained action network stabilize the tracking error system. Finally, the modified Van der Pol's oscillator system is employed to implement the Q-learning algorithm and verify the theoretical results.

REFERENCES

- [1] D. Liu, S. Xue, B. Zhao, B. Luo, and Q. Wei, "Adaptive dynamic programming for control: A survey and recent advances," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 142–160, Jan. 2021.
- [2] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming with Applications in Optimal Control*. Cham, Switzerland: Springer, 2017.
- [3] Y. Yan, D. Wang, H. Li, and D. Liu, "Error bound analysis of Q-function for discounted optimal control problems with policy iteration," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1207–1216, July 2017.
- [4] Z. Liang, M. Ha, and D. Liu, "Theoretical analysis of value-iteration-based Q-learning with approximation errors," in *12th International Conference on Information Science and Technology (ICIST)*, Kaifeng, China, 2022, pp. 120–125.
- [5] Y. -M. Park, M. -S. Choi, and K. Y. Lee, "An optimal tracking neuro-controller for nonlinear dynamic systems," *IEEE Transactions on Neural Networks*, vol. 7, no. 5, pp. 1099–1110, Sept. 1996.
- [6] H. Zhang, R. Song, Q. Wei, and T. Zhang, "Optimal tracking control for a class of nonlinear discrete-time systems with time delays based on heuristic dynamic programming," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 1851–1862, Dec. 2011.
- [7] B. Kiumarsi and F. L. Lewis, "Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 1, pp. 140–151, Jan. 2015.
- [8] B. Luo, D. Liu, T. Huang, and D. Wang, "Model-free optimal tracking control via critic-only Q-learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 2134–2144, Oct. 2016.
- [9] C. Li, J. Ding, F. L. Lewis, and T. Chai, "A novel adaptive dynamic programming based on tracking error for nonlinear discrete-time systems," *Automatica*, vol. 129, no. 109687, May 2021.
- [10] M. Ha, D. Wang, and D. Liu, "Discounted iterative adaptive critic designs with novel stability analysis for tracking control," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1262–1272, July 2022.
- [11] R. Song and L. Zhu, "Optimal fixed-point tracking control for discrete-time nonlinear systems via ADP," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 657–666, May 2019.
- [12] C. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 1989.
- [13] C. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992.
- [14] Q. Wei, D. Liu, Q. Lin, and R. Song, "Adaptive dynamic programming for discrete-time zero-sum games," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 957–969, Apr. 2018.