IET Control Theory & Applications

Research Article

IET Journals
The Institution of Engineering and Technology

Using reinforcement learning techniques to solve continuous-time non-linear optimal tracking problem without system dynamics

ISSN 1751-8644
Received on 6th August 2015
Revised on 10th November 2015
Accepted on 1st December 2015
doi: 10.1049/iet-cta.2015.0769
www.ietdl.org

Yuanheng Zhu[1], Dongbin Zhao[1, ✉], Xiangjun Li[2]

[1]State Key Laboratory of Management and Control for Complex Systems, Institution of Automation, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
[2]Electrical Engineering and New Material Department, China Electric Power Research Institute, Beijing 100192, People's Republic of China
✉ E-mail: dongbin.zhao@ia.ac.cn

Abstract: The optimal tracking of non-linear systems without knowing system dynamics is an important and intractable problem. Based on the framework of reinforcement learning (RL) and adaptive dynamic programming, a model-free adaptive optimal tracking algorithm is proposed in this study. After constructing an augmented system with the tracking errors and the reference states, the tracking problem is converted to a regulation problem with respect to the new system. Several RL techniques are synthesised to form a novel algorithm which learns the optimal solution online in real time without any information of the system dynamics. Continuous adaptation laws are defined by the current observations and the past experience. The convergence is guaranteed by Lyapunov analysis. Two simulations on a linear and a non-linear systems demonstrate the performance of the proposed approach.

1 Introduction

In the field of control theory and its applications, reinforcement learning (RL) [1, 2] and adaptive dynamic programming (ADP) [3, 4] have been intensively investigated as a forward-in-time solution to the optimal control problems. In the past few years, many works concentrated on optimal regulation problems with RL/ADP algorithms [5–10]. The target is to adaptively learn a controller that balances the system costs and the control efforts in an optimal manner. These studies are meaningful both from the theoretical viewpoint and for applicatory values. Surveys about RL and ADP are available in [11, 12].

RL and ADP researchers also pay sufficient attention to another control issue which is known as the optimal tracking problem [13–16]. In many practical situations, systems are expected to track certain reference trajectories. Such demand is widely confronted in underactuated vehicles, flight control and power systems. The optimal tracking objective is to minimise the difference between the controlled system and the reference trajectories with the control efforts as small as possible. Conventional approaches via RL/ADP consist of a feedforward part and a feedback part [17–22]. The feedforward controller computes the desired actions to maintain the desired trajectories, while the feedback controller aims to eliminate the tracking errors based on RL and ADP. Such design has its limitations. The knowledge of the reference dynamics is indispensable to design the feedforward controller, and the system shall satisfy certain dynamics inversion condition which is unpractical for most systems.

To overcome the limited design, a novel scheme was proposed in [23, 24]. The main principle is to define an augmented system that integrates both the tracking errors and the reference states into one synthesised dynamics. The control objective thus becomes finding a regulator that stabilises the error states of the augmented system in the optimal sense. The feedforward and feedback controllers are merged together in the scheme, and the tracking problem is converted to a regulation problem, which makes it easier to apply the state-of-the-art RL techniques.

RL is first proposed in computational intelligence community mainly to solve discrete-time Markov decision problems. The extension to continuous-time (CT) systems is followed by the development of ADP. In [6], Abu-Khalaf and Lewis apply policy iteration (PI) to the optimal control of CT non-linear systems with saturating controllers. To avoid dealing with the time derivatives in CT dynamics, Vrabie and Lewis [25] come up with the integral RL (IRL) technique. IRL benefits their algorithm in the fact that the knowledge of the internal dynamics is no longer needed. In [26], Vamvoudakis *et al.* employ actor–critic structure in their synchronous PI algorithm. Both values and policies are approximated by neural networks (NNs) and are tuned simultaneously using online observations. To address problems with completely unknown dynamics, Jiang and Jiang [27] consider CT linear systems and give a model-free approach that iteratively learns the optimal solutions based on the system observations. In their algorithm, the observations are produced by a policy that differs to the evaluated one. This process corresponds to the idea of off-policy in the RL literatures, which refers to the fact that the executed policies are different to the estimated ones. Such mechanism is extended to non-linear systems in [28, 29]. Another intelligent technique in the field of RL, experience replay (ER), is successfully applied by Modares *et al.* [30] in their adaptive algorithm. Past observations are repeatedly utilised and the learning speed is improved dramatically.

The rapid development of ADP promotes the research in optimal tracking control. In [31], Modare and Lewis combine PI with IRL to solve the CT linear optimal tracking problem. Li *et al.* [32] consider the finite horizon cost function and convert the problem to the solution of a time-varying Riccati equation. To address the tracking problem in an online real-time manner, Modares and Lewis [23] extend synchronous PI algorithm to CT non-linear systems and establish the stability property of the whole system. Note that the above mentioned works require the system dynamics at least partially known. The input gain matrix is necessary when updating policies. To break the limitation, inspired by Jiang and Jiang [27], Qin *et al.* [33] present a completely model-free algorithm to the CT linear optimal tracking problem. Furthermore, Modares *et al.* [34] investigate the tracking control of completely unknown non-linear systems via off-policy RL technique. The disturbance is considered in the system and H_∞ control is used to attenuate the negative effect. After a group of observations are collected, offline iterations proceed by updating policies on the basis of the

last calculated ones. Since the data is collected beforehand, the learning is not sensitive to the real-time process. Motivated by that, a novel algorithm is presented in this paper.

We consider the system with the CT non-linear dynamics and propose an online model-free adaptive optimal tracking algorithm that learns an optimal tracker in a real-time manner. First the tracking problem is converted to an augmented system that is composed of the tracking errors and the reference states. The optimal tracking control is equivalent to the optimal regulation solution with respect to the new dynamics. Several advanced RL techniques are utilised to solve the problem, such as PI, IRL, actor–critic and ER. The convergence is established by the Lyapunov analysis. The novel algorithm needs neither the controlled dynamics nor the reference model, and updates critic and actor NNs with continuous adaptations. A linear and a non-linear experiments are simulated. The results are consistent with our theoretical conclusions.

2 Problem formulation

The control objective herein is to drive the outputs of a system to track certain reference trajectories. The controlled system is described by the CT non-linear input-affine dynamics

$$\begin{aligned} \dot{x} &= f(x) + g(x)u \\ y &= Cx \end{aligned} \tag{1}$$

where the state $x \in \mathbb{R}^n$, the control $u \in \mathbb{R}^m$ and the output $y \in \mathbb{R}^l$. The drift dynamics $f \in \mathbb{R}^n$ and the input gain matrix $g \in \mathbb{R}^{n\times m}$ are Lipschitz continuous and $f(0) = 0$. $C \in \mathbb{R}^{l\times n}$ represents the output matrix. It is assumed that the system is controllable on a compact set $\Omega \in \mathbb{R}^n$.

The reference trajectories are produced by a command generator with the dynamics

$$\begin{aligned} \dot{x}_\mathrm{d} &= h(x_\mathrm{d}) \\ y_\mathrm{d} &= Cx_\mathrm{d} \end{aligned} \tag{2}$$

where $x_\mathrm{d} \in \mathbb{R}^n$ and the dynamics function $h \in \mathbb{R}^n$ is Lipschitz continuous. The target is to find a controller such that y can track y_d with the minimum tracking error $e_y = y - y_\mathrm{d}$. Throughout this paper, it is assumed that the command generator is stable on Ω, not necessarily to be asymptotically stable.

Given a tracker $u(t)$, its tracking performance is determined by an infinite horizon discounted cost index J

$$\begin{aligned} J(x(t), x_\mathrm{d}, u) &= \int_t^\infty e^{-\gamma(\tau - t)} \left[(y - y_\mathrm{d})^\mathrm{T} Q_\mathrm{e} (y - y_\mathrm{d}) + u^\mathrm{T} R u \right] \mathrm{d}\tau \\ &= \int_t^\infty e^{-\gamma(\tau - t)} \left[(x - x_\mathrm{d})^\mathrm{T} C^\mathrm{T} Q_\mathrm{e} C (x - x_\mathrm{d}) + u^\mathrm{T} R u \right] \mathrm{d}\tau \end{aligned}$$

where $Q_\mathrm{e} \in \mathbb{R}^{l\times l}$ is a positive definite matrix and $R \in \mathbb{R}^{m\times m}$ is a positive symmetric matrix. $\gamma > 0$ is the discounted factor which is necessary to guarantee the index valid since the control input may keep varying to track the reference trajectories.

To solve the optimal tracking problem, an augmented dynamics is defined with the tracking error $e_x = x - x_\mathrm{d}$ and the reference state x_d. Based on (1) and (2)

$$\begin{aligned} \dot{e}_x &= \dot{x} - \dot{x}_\mathrm{d} \\ &= f(e_x + x_\mathrm{d}) + g(e_x + x_\mathrm{d})u - h(x_\mathrm{d}) \end{aligned}$$

Let $X = [e_x^\mathrm{T}, x_\mathrm{d}^\mathrm{T}]^\mathrm{T}$, and the augmented dynamics has

$$\begin{aligned} \dot{X} &= \begin{bmatrix} f(e_x + x_\mathrm{d}) - h(x_\mathrm{d}) \\ h(x_\mathrm{d}) \end{bmatrix} + \begin{bmatrix} g(e_x + x_\mathrm{d}) \\ 0 \end{bmatrix} u \\ &= F(X) + G(X)u \end{aligned} \tag{3}$$

As a result, the value of a tracker u w.r.t. (3) on Ω becomes

$$V(x(t)) = \int_t^\infty e^{-\gamma(\tau - t)} \left[X^\mathrm{T} Q X + u^\mathrm{T} R u \right] \mathrm{d}\tau \tag{4}$$

where

$$Q = \begin{bmatrix} C^\mathrm{T} Q_\mathrm{e} C & 0 \\ 0 & 0 \end{bmatrix} \geq 0.$$

Definition 1 Admissible [23]: Given a state feedback policy $u(t) = u(X(t)) \in \mathbb{R}^m$. If u is continuous on Ω, $u(0) = 0$, u stabilises (3) and $V(X)$ is finite $\forall X \in \Omega$, we say u is an admissible policy w.r.t. (3) on Ω, denoted by $u \in \psi(\Omega)$.

Now the problem becomes finding the optimal admissible policy u^* that has the minimum value, i.e. the optimal value function V^*. An infinitesimal version of (4) is the Bellman equation

$$\nabla V^\mathrm{T}(F + Gu) - \gamma V + X^\mathrm{T} Q X + u^\mathrm{T} R u = 0, \quad V(0) = 0 \tag{5}$$

Define the Hamiltonian function

$$H(X, V, u) = \nabla V^\mathrm{T}(F + Gu) - \gamma V + X^\mathrm{T} Q X + u^\mathrm{T} R u$$

According to the stationary condition, the optimal tracker u^* is derived

$$\frac{\partial H}{\partial u}(X, V^*, u) = 0 \Rightarrow u^*(X) = -\frac{1}{2} R^{-1} G^\mathrm{T}(X) \nabla V^*(X) \tag{6}$$

After inserting into (5), the *Hamilton–Jacobi–Bellman* (HJB) equation is obtained

$$\nabla V^{*\mathrm{T}} F - \frac{1}{4} \nabla V^{*\mathrm{T}} G R^{-1} G^\mathrm{T} \nabla V^* - \gamma V^* + X^\mathrm{T} Q X = 0, V^*(0) = 0 \tag{7}$$

In [23], Modares and Lewis have pointed out that in the limit as γ approaches 0, with the smooth positive definite solution V^*, the corresponding optimal tracker u^* asymptotically stabilises the tracking error. Unfortunately in practical situations, the reference trajectories usually do not go to zero, so γ has to be greater than 0 to make V valid. However, the authors also suggest if we select smaller γ and/or define larger Q_e, the boundness of the tracking error can be constrained as small as desired.

To determine the optimal tracker, one needs to solve the HJB equation which apparently is a non-linear partial differential equation. A computational approach is *policy iteration*. Given an initial admissible policy u_0, iterate the following two steps until the final convergence is achieved.

(i) (*Policy evaluation*) Solve the Bellman equation

$$\nabla V_i^\mathrm{T}(F + Gu_i) - \gamma V_i + X^\mathrm{T} Q X + u_i^\mathrm{T} \ Ru_i = 0, \ \ V_i(0) = 0 \tag{8}$$

(ii) (*Policy improvement*) Produce an improved policy

$$u_{i+1}(X) = -\frac{1}{2} R^{-1} G^\mathrm{T}(X) \nabla V_i(X) \tag{9}$$

It is easy to prove that V_i and u_i converge to V^* and u^* as $i \to \infty$. However, in the sight of the above PI equations, (8) requires the complete knowledge of the augmented system dynamics. By applying the IRL technique we can avoid the dependence on the drift dynamics f, but the input gain matrix g is still necessary in (9). In various situations, both the controlled model and the reference dynamics are unknown to the designers. An additional model identifier will increase the computational cost and decrease the calculation accuracy. It is more desirable to devise an online adaptive algorithm for this problem. In the following sections, we address this issue and propose a model-free online adaptive approach to the optimal tracking problem without any system dynamics.

3 Adaptive optimal tracking learning with completely unknown dynamics

Consider an arbitrary control effort u is applied to (3). Suppose an admissible policy u_0 is already known, and one desires to determine the value V_1 and the improved policy u_1. Differentiate the item $e^{-\gamma t}V_0(t)$ along the system solutions $(F+Gu)$

$$\left(e^{-\gamma t}V_0(t)\right)' = -\gamma e^{-\gamma t}V_0 + e^{-\gamma t}\nabla V_0^{\mathrm{T}}(F+Gu)$$

For the ease of expression, the variable X is dropped in the above equation as well as in the sequel. According to (8) and (9), it is further inferred that

$$\begin{aligned}\left(e^{-\gamma t}V_0(t)\right)' &= e^{-\gamma t}\left[-\gamma V_0 + \nabla V_0^{\mathrm{T}}(F+Gu_0) + \nabla V_0^{\mathrm{T}}G(u-u_0)\right]\\ &= e^{-\gamma t}\left[-X^{\mathrm{T}}QX - u_0^{\mathrm{T}}Ru_0 - 2u_1^{\mathrm{T}}R(u-u_0)\right]\end{aligned}$$

Based on the *integral reinforcement learning* [25], a novel Bellman equation is obtained with some manipulations after integrating the above equation on both sides along the interval $[t-T,t]$

$$\begin{aligned}V_0(t) - e^{\gamma \mathrm{T}}V_0(t-T) + \int_{t-T}^{t} e^{-\gamma(\tau-t)}\\ \times\left[X^{\mathrm{T}}QX + u_0^{\mathrm{T}}Ru_0 + 2u_1^{\mathrm{T}}R(u-u_0)\right]\mathrm{d}\tau = 0\end{aligned} \tag{10}$$

Compared to (8), the new equation contains no dynamics of the augmented system and both V_0 and u_1 are calculated by solving only one equation. The following theorem demonstrates the equivalence between (10) and (8), (9).

Theorem 1: Let $\bar{t}_j \in [t-T,t]$ $(j=0,1,\ldots,L)$ be the time instants satisfying

$$\bar{t}_0 = t-T \le \bar{t}_1 \le \bar{t}_2 \le \cdots \le \bar{t}_L = t$$

Let $e_u = u - u_0$ and assume that e_u is piecewise constant and determined by $e_u(\tau) = c_j, \forall \tau \in [\bar{t}_j, \bar{t}_{j+1})$, where $\{c_j\}_{j=1}^{L}$ is a sequence of constant vectors in $\mathbb{R}^m$. Besides, there exist $\beta_1, \beta_2 > 0$ such that

$$\beta_1 I \le \sum_{j=1}^{L-1}(c_j - c_{j+1})(c_j - c_{j+1})^{\mathrm{T}} \le \beta_2 I \tag{11}$$

where I denotes the identity matrix with the appropriate dimension. Then the solutions to (10) are uniquely determined by (8) and (9).

Proof: The theorem is proved in the same way as [35], so it is omitted here. □

If we let $u_0 = u^*$, then $V_1 = V^*$ and $u_1 = u^*$. Consequently, the Bellman equation (10) is converted to a model-free HJB equation

$$\begin{aligned}V^*(t) - e^{\gamma \mathrm{T}}V^*(t-T) + \int_{t-T}^{t} e^{-\gamma(\tau-t)}\\ \times\left[X^{\mathrm{T}}QX + 2u^{*\mathrm{T}}Ru - u^{*\mathrm{T}}Ru^*\right]\mathrm{d}\tau = 0\end{aligned} \tag{12}$$

The optimal tracking problem now becomes solving (12) for V^* and u^*.

Remark 1: In the theorem, the probing noise e_u is required to satisfy (11), which can be seen as a *persistency of excitation* (PE) condition in adaptive control. To meet the arguments in the proof, e_u is restricted to be piecewise constant in contrast to the commonly used random noises or sinusoid signals. In our simulated experiments, we select sinusoidal signals as the probing noise and the method can still learn the optimal tracking solutions.

4 Model-free adaptive optimal tracking algorithm

4.1 NN approximation and approximate solutions to the Bellman equation

Suppose we are given an admissible u_0 and try to determine V_0 and u_1 by solving (10). To efficiently express the value and policy functions, we employ the universal approximation property of NNs. According to the Weirstrass high-order approximation theorem, the smooth value function V_0 can be uniformly approximated over a compact set by a set of linearly independent basis functions

$$V_0(X) = W_{\mathrm{c}}^{\mathrm{T}}\phi_{\mathrm{c}}(X) + \varepsilon_{\mathrm{c}} \tag{13}$$

where $W_{\mathrm{c}} \in \mathbb{R}^{K_1}$ represent the ideal coefficients, $\phi_{\mathrm{c}} \in \mathbb{R}^{K_1}$ is the basis function vector and $\varepsilon_{\mathrm{c}} \in \mathbb{R}$ is the approximation error. K_1 indicates the number of neurons in the hidden layer. As $K_1 \to \infty$, $\varepsilon_{\mathrm{c}} \to 0$. Similarly, a smooth policy u_1 can be expressed by

$$u_1(X) = W_{\mathrm{a}}^{\mathrm{T}}\phi_{\mathrm{a}}(X) + \varepsilon_{\mathrm{a}} \tag{14}$$

where $W_{\mathrm{a}} \in \mathbb{R}^{K_2\times m}, \phi_{\mathrm{a}} \in \mathbb{R}^{K_2}, \varepsilon_{\mathrm{a}} \in \mathbb{R}^m$ and K_2 denotes the number of hidden neurons.

Assumption 1: Suppose the control effort u is a stabilisable input such that the closed-loop system $(F+Gu)$ remains in the compact set Ω for any starting state $X(0) \in \Omega$.

Assumption 2: (a) The ideal coefficients W_{c} and W_{a} are bounded so that

$$\|W_{\mathrm{c}}\| \le W_{c\max}$$
$$\|W_{\mathrm{a}}\| \le W_{a\max}$$

(b) The basis function vectors ϕ_{c} and ϕ_{a} are bounded so that

$$\|\phi_{\mathrm{c}}\| \le b_{\phi c}$$
$$\|\phi_{\mathrm{a}}\| \le b_{\phi a}$$

(c) The approximation errors ε_{c} and ε_{a} are bounded so that

$$\|\varepsilon_{\mathrm{c}}\| \le b_{\varepsilon c}$$
$$\|\varepsilon_{\mathrm{a}}\| \le b_{\varepsilon a}$$

Suppose V_0 and u_1 are represented in the form of (13) and (14). After inserting them into (10) we yield the approximate Bellman equation

$$\begin{aligned}\varepsilon_{\mathrm{B}} = W_{\mathrm{c}}^{\mathrm{T}}\left[\phi_{\mathrm{c}}(t) - e^{\gamma \mathrm{T}}\phi_{\mathrm{c}}(t-T)\right] + \int_{t-T}^{t} e^{-\gamma(\tau-t)}\\ \times\left[X^{\mathrm{T}}QX + u_0^{\mathrm{T}}Ru_0 + 2\phi_{\mathrm{a}}^{\mathrm{T}}W_{\mathrm{a}}R(u-u_0)\right]\mathrm{d}\tau\end{aligned} \tag{15}$$

where ε_{B} is caused by approximation errors and has

$$\varepsilon_{\mathrm{B}} = -\varepsilon_{\mathrm{c}}(t) + e^{\gamma \mathrm{T}}\varepsilon_{\mathrm{c}}(t-T) - \int_{t-T}^{t} 2e^{-\gamma(\tau-t)}\varepsilon_{\mathrm{a}}^{\mathrm{T}}R(u-u_0)\,\mathrm{d}\tau$$

Since approximation errors are bounded, ε_{B} is also bounded, denoted by $\|\varepsilon_{\mathrm{B}}\| \le b_{\mathrm{B}}$.

To determine the ideal values of W_{c} and W_{a}, define the estimations $\hat{W}_{\mathrm{c}}$ and $\hat{W}_{\mathrm{a}}$, and adopt the *actor–critic* structure to

approximate V_0 and u_1 by

$$\hat{V}_0 = \hat{W}_c^T \phi_c$$
$$\hat{u}_1 = \hat{W}_a^T \phi_a$$

Consequently, a residual error is formulated based on the Kronecker product representation

$$e_1 = \hat{W}_c^T \left[\phi_c(t) - e^{\gamma T}\phi_c(t-T)\right] + \mathbf{v}(\hat{W}_a)^T \times \int_{t-T}^{t} 2e^{-\gamma(\tau-t)} \left[R(u-u_0)\otimes\phi_a\right] d\tau + \int_{t-T}^{t} e^{-\gamma(\tau-t)} \left(X^T Q X + u_0^T R u_0\right) d\tau \quad (16)$$

where $\otimes$ represents the Kronecker product and $\mathbf{v}(\cdot)$ is the vectorising operator that transforms a matrix to a vector by stacking the columns one by one. If we define

$$\sigma_1(t) = \phi_c(t) - e^{\gamma T}\phi_c(t-T)$$
$$\eta_1(t) = \int_{t-T}^{t} 2e^{-\gamma(\tau-t)} \left[R(u-u_0)\otimes\phi_a\right] d\tau$$
$$p_1(t) = \int_{t-T}^{t} e^{-\gamma(\tau-t)} \left(X^T Q X + u_0^T R u_0\right) d\tau$$

and let

$$\hat{Z} = \begin{bmatrix} \hat{W}_c \\ \mathbf{v}(\hat{W}_a) \end{bmatrix} \quad \rho_1(t) = \begin{bmatrix} \sigma_1(t) \\ \eta_1(t) \end{bmatrix},$$

e_1 becomes

$$e_1(t) = \hat{Z}^T \rho_1(t) + p_1(t) \quad (17)$$

So the gradient descent method is utilisable to train $\hat{Z}$ along the gradient $\partial e_1/\partial \hat{Z}$ in order to minimise $(1/2)\, e_1^2$.

Reviewing (16), the past system data are also exploitable to define other residual errors with the current estimations $\hat{W}_c$ and $\hat{W}_a$. Let

$$\sigma_1(t_j) = \phi_c(t_j) - e^{\gamma T}\phi_c(t_j - T)$$
$$\eta_1(t_j) = \int_{t_j-T}^{t_j} 2e^{-\gamma(\tau-t_j)} \left[R(u-u_0)\otimes\phi_a\right] d\tau$$
$$p_1(t_j) = \int_{t_j-T}^{t_j} e^{-\gamma(\tau-t_j)} (X^T Q X + u_0^T R u_0)\, d\tau$$

and define

$$e_1(t_j) = \hat{Z}^T \rho_1(t_j) + p_1(t_j)$$

where t_j represents the past instant and

$$\rho_1(t_j) = \begin{bmatrix} \sigma_1(t_j) \\ \eta_1(t_j) \end{bmatrix}.$$

So we can define a more comprehensive adaptation law with the target to minimise the sum of a group of residual errors defined not only by the current data, but also by the past experience. Such design corresponds to the idea of *experience replay* in the field of RL [1] and is helpful to improve the learning rate and accuracy.

Suppose a set of past experiences are stored, termed as $\left\{\left(\sigma(t_j), \eta_1(t_j), p_1(t_j)\right)\right\}_{j=1}^{N}$. The adaptation law for the critic and actor weights using ER is described by

$$\dot{\hat{Z}} = \begin{bmatrix} \dot{\hat{W}}_c \\ \dot{\hat{W}}_a \end{bmatrix} = -\frac{\alpha}{N+1} \left\{ \frac{\rho_1(t)}{m_1^2(t)} \left[\hat{Z}^T \rho_1(t) + p_1(t)\right] + \sum_{j=1}^{N} \frac{\rho_1(t_j)}{m_1^2(t_j)} \left[\hat{Z}^T \rho_1(t_j) + p_1(t_j)\right] \right\} \quad (18)$$

where $m_1 = \left(\rho_1^T \rho_1 + 1\right)^{\frac{1}{2}}$ is for normalisation and α is the learning rate. The following theorem demonstrates the convergence of the adaptation. Before the theorem, a persistency of excitation condition is introduced first.

Assumption 3: For any time interval $[t-T, t]$, there exist two constants β_3 and β_4 such that the signal $\bar{\rho}_1 = \rho_1/(\rho_1^T \rho_1 + 1)^{\frac{1}{2}}$ always has

$$\beta_3 I \le \bar{\rho}_1(t)\bar{\rho}_1^T(t) \le \beta_4 I$$

Theorem 2: Given an admissible u_0, let $Z = [W_c^T, W_a^T]^T$ where W_c and W_a are the ideal coefficients of V_0 and u_1. Let $\hat{Z} = [\hat{W}_c^T, \hat{W}_a^T]^T$ be the estimations of the critic and actor NNs. Suppose Assumptions 1–3 hold and $\hat{Z}$ are tuned following (18). Then $\hat{Z}$ converge exponentially to a residual set in the neighbour of the ideal Z.

Proof: Define the estimation error $\tilde{Z} = Z - \hat{Z}$ and the Lyapunov candidate

$$L_1 = \frac{1}{2}\tilde{Z}^T \alpha^{-1} \tilde{Z}$$

Its time derivative has

$$\dot{L}_1 = \tilde{Z}^T \alpha^{-1} \dot{\tilde{Z}}$$

From (15) and (17), we have

$$\varepsilon_B = Z^T \rho_1 + p_1$$
$$e_1 = \varepsilon_B - \tilde{Z}^T \rho_1$$

So the adaptation law (18) can be rewritten as

$$\dot{\hat{Z}} = -\frac{\alpha}{N+1} \left\{ \frac{\rho_1(t)}{m_1^2(t)} \left[\varepsilon_B(t) - \tilde{Z}^T \rho_1(t)\right] + \sum_{j=1}^{N} \frac{\rho_1(t_j)}{m_1^2(t_j)} \left[\varepsilon_B(t_j) - \tilde{Z}^T \rho_1(t_j)\right] \right\}$$

Utilising $\dot{\tilde{Z}} = -\dot{\hat{Z}}$, $\dot{L}_1$ becomes

$$\dot{L}_1 = \frac{1}{N+1} \left\{ -\tilde{Z}^T \left[\bar{\rho}_1(t)\bar{\rho}_1^T(t) + \sum_{j=1}^{N} \bar{\rho}_1(t_j)\bar{\rho}_1^T(t_j)\right] \tilde{Z} + \left[\varepsilon_B(t)\frac{\bar{\rho}_1^T(t)}{m_1(t)} + \sum_{j=1}^{N} \varepsilon_B(t_j)\frac{\bar{\rho}_1^T(t_j)}{m_1(t_j)}\right] \tilde{Z} \right\}$$
$$\le -\frac{1}{N+1}\lambda_{\min}(\bar{H}_1) \left\|\tilde{Z}\right\|^2 + b_B \left\|\tilde{Z}\right\|$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of the given matrix and

$$\bar{H}_1 = \bar{\rho}_1(t)\bar{\rho}_1^T(t) + \sum_{j=1}^{N} \bar{\rho}_1(t_j)\bar{\rho}_1^T(t_j)$$

Under the PE condition, $\tilde{Z}$ converge to the residual set

$$\left\{\tilde{Z} : \left\|\tilde{Z}\right\| \le \frac{(N+1)b_B}{\lambda_{\min}(\bar{H}_1)}\right\} \quad (19)$$

□

From the above analysis, the convergence rate is determined by $\lambda_{\min}(\bar{H}_1)$, which can be improved by increasing the minimum eigenvalue of $H_1 = \sum_{j=1}^{N} \bar{\rho}_1(t_j)\bar{\rho}_1^{\mathrm{T}}(t_j)$. During the learning process, the new observation replaces some old data in the experience set only if such replacement results in an increase of $\lambda_{\min}(H_1)$. In addition, increasing $\lambda_{\min}(H_1)$ helps to reduce the estimation error according to the results given in (19).

4.2 Adaptive optimal solution to the optimal tracking problem

We extend the supposition that the optimal V^* and u^* are also represented in the forms (13) and (14), respectively. After inserting into (12), we obtain

$$\varepsilon_H = W_{\mathrm{c}}^{\mathrm{T}}\left[\phi_{\mathrm{c}}(t) - e^{\gamma \mathrm{T}}\phi_{\mathrm{c}}(t-T)\right] + \int_{t-T}^{t} e^{-\gamma(\tau-t)} \times \left[X^{\mathrm{T}}QX + 2\phi_{\mathrm{a}}^{\mathrm{T}}W_{\mathrm{a}}Ru - \phi_{\mathrm{a}}^{\mathrm{T}}W_{\mathrm{a}}RW_{\mathrm{a}}^{\mathrm{T}}\phi_{\mathrm{a}}\right] \mathrm{d}\tau \tag{20}$$

where

$$\varepsilon_H = -\varepsilon_{\mathrm{c}}(t) + e^{\gamma \mathrm{T}}\varepsilon_{\mathrm{c}}(t-T) - \int_{t-T}^{t} e^{-\gamma(\tau-t)} \times \left[2\varepsilon_{\mathrm{a}}^{\mathrm{T}}Ru - 2\varepsilon_{\mathrm{a}}^{\mathrm{T}}RW_{\mathrm{a}}^{\mathrm{T}}\phi_{\mathrm{a}} - \varepsilon_{\mathrm{a}}^{\mathrm{T}}R\varepsilon_{\mathrm{a}}\right] \mathrm{d}\tau$$

Since the approximation errors and basis functions are bounded, ε_H is bounded by $\|\varepsilon_H\| \le b_H$.

Critic and actor NNs are constructed to approximate V^* and u^*

$$\hat{V}^*(X) = \hat{W}_{\mathrm{c}}^{\mathrm{T}}\phi_{\mathrm{c}}(X)$$
$$\hat{u}^*(X) = \hat{W}_{\mathrm{a}}^{\mathrm{T}}\phi_{\mathrm{a}}(X)$$

An HJB error is obtained consequently

$$e_2 = \hat{W}_{\mathrm{c}}^{\mathrm{T}}\left[\phi_{\mathrm{c}}(t) - e^{\gamma \mathrm{T}}\phi_{\mathrm{c}}(t-T)\right] + \int_{t-T}^{t} e^{-\gamma(\tau-t)} \times \left[X^{\mathrm{T}}QX + 2\phi_{\mathrm{a}}^{\mathrm{T}}\hat{W}_{\mathrm{a}}Ru - \phi_{\mathrm{a}}^{\mathrm{T}}\hat{W}_{\mathrm{a}}R\hat{W}_{\mathrm{a}}^{\mathrm{T}}\phi_{\mathrm{a}}\right] \mathrm{d}\tau \tag{21}$$

Let

$$\begin{aligned}
\sigma_2(t) &= \phi_{\mathrm{c}}(t) - e^{\gamma \mathrm{T}}\phi_{\mathrm{c}}(t-T)\\
\mu_2(t) &= \int_{t-T}^{t} e^{-\gamma(\tau-t)}\,(Ru \otimes \phi_{\mathrm{a}})\,\mathrm{d}\tau\\
D_2(t) &= \int_{t-T}^{t} e^{-\gamma(\tau-t)}\left(R \otimes \phi_{\mathrm{a}}\phi_{\mathrm{a}}^{\mathrm{T}}\right)\mathrm{d}\tau\\
p_2(t) &= \int_{t-T}^{t} e^{-\gamma(\tau-t)}\left(X^{\mathrm{T}}QX\right)\mathrm{d}\tau
\end{aligned} \tag{22}$$

and to facilitate the derivation of the adaptation law, define

$$\eta_2(t) = 2\mu_2(t) - 2D_2(t)\mathbf{v}(\hat{W}_{\mathrm{a}})$$

So e_2 is rewritten as

$$e_2(t) = \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma_2(t) + \mathbf{v}(\hat{W}_{\mathrm{a}})^{\mathrm{T}}\eta_2(t) + \mathbf{v}(\hat{W}_{\mathrm{a}})^{\mathrm{T}}D_2(t)\mathbf{v}(\hat{W}_{\mathrm{a}}) + p_2(t)$$

or

$$e_2(t) = \hat{Z}^{\mathrm{T}}\rho_2(t) + \mathbf{v}(\hat{W}_{\mathrm{a}})^{\mathrm{T}}D_2(t)\mathbf{v}(\hat{W}_{\mathrm{a}}) + p_2(t)$$

where

$$\hat{Z} = \begin{bmatrix}\hat{W}_{\mathrm{c}}\\ \mathbf{v}(\hat{W}_{\mathrm{a}})\end{bmatrix} \quad \text{and} \quad \rho_2(t) = \begin{bmatrix}\sigma_2(t)\\ \eta_2(t)\end{bmatrix}.$$

To apply the ER technique, a set of experience data are stored in $\left\{\left(\sigma_2(t_j), \mu_2(t_j), D_2(t_j), p_2(t_j)\right)\right\}_{j=1}^{N}$, whose elements are defined in the same way as (22), but with time indices $\{t_j\}_{j=1}^{N}$. Meanwhile, we define

$$\begin{aligned}
\eta_2(t_j) &= 2\mu_2(t_j) - 2D_2(t_j)\mathbf{v}(\hat{W}_{\mathrm{a}})\\
\rho_2(t_j) &= \begin{bmatrix}\sigma_2(t_j)\\ \eta_2(t_j)\end{bmatrix}\\
e_2(t_j) &= \hat{Z}^{\mathrm{T}}\rho_2(t_j) + \mathbf{v}(\hat{W}_{\mathrm{a}})^{\mathrm{T}}D_2(t_j)\mathbf{v}(\hat{W}_{\mathrm{a}}) + p_2(t_j)
\end{aligned}$$

The gradient-based adaptation law for $\hat{Z}$ is given by

$$\begin{aligned}
\dot{\hat{Z}} &= \begin{bmatrix}\dot{\hat{W}}_{\mathrm{c}}\\ \dot{\hat{W}}_{\mathrm{a}}\end{bmatrix}\\
&= -\frac{\alpha}{N+1}\left\{\frac{\rho_2(t)}{m_2^2(t)}\left[\hat{Z}^{\mathrm{T}}\rho_2(t) + \mathbf{v}(\hat{W}_{\mathrm{a}})^{\mathrm{T}}D_2(t)\mathbf{v}(\hat{W}_{\mathrm{a}}) + p_2(t)\right]\right.\\
&\quad \left. + \sum_{j=1}^{N}\frac{\rho_2(t_j)}{m_2^2(t_j)}\left[\hat{Z}^{\mathrm{T}}\rho_2(t_j) + \mathbf{v}(\hat{W}_{\mathrm{a}})^{\mathrm{T}}D_2(t_j)\mathbf{v}(\hat{W}_{\mathrm{a}}) + p_2(t_j)\right]\right\}
\end{aligned} \tag{23}$$

where $m_2 = \left(\rho_2^{\mathrm{T}}\rho_2 + 1\right)^{\frac{1}{2}}$.

Theorem 3: Let $Z = [W_{\mathrm{c}}^{\mathrm{T}}, W_{\mathrm{a}}^{\mathrm{T}}]^{\mathrm{T}}$ be the ideal NN coefficients for V^* and u^*. Let $\hat{Z} = [\hat{W}_{\mathrm{c}}, \hat{W}_{\mathrm{a}}]^{\mathrm{T}}$ be the estimations in the critic and actor. Assume the signal $\bar{\rho}_2 = \rho_2/(\rho_2^{\mathrm{T}}\rho_2 + 1)^{\frac{1}{2}}$ is persistently exciting. Under Assumptions 1, 2 and the tuning law provided by (23), $\hat{Z}$ converge to a residual set in the neighbour of Z.

Proof: Define the estimation error $\tilde{Z} = Z - \hat{Z}$ and the Lyapunov candidate

$$L_2 = \frac{1}{2}\tilde{Z}^{\mathrm{T}}\alpha^{-1}\tilde{Z}$$

After adding and subtracting both sides of (20) into (21), we obtain

$$\begin{aligned}
e_2 &= \varepsilon_H - \tilde{W}_{\mathrm{c}}^{\mathrm{T}}\sigma_2 - 2\mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}\mu_2 + 2\mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2\mathbf{v}(W_{\mathrm{a}})\\
&\quad - \mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2\mathbf{v}(\tilde{W}_{\mathrm{a}})\\
&= \varepsilon_H - \tilde{W}_{\mathrm{c}}^{\mathrm{T}}\sigma_2 - 2\mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}\mu_2 + 2\mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2\mathbf{v}(\hat{W}_{\mathrm{a}})\\
&\quad + \mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2\mathbf{v}(\tilde{W}_{\mathrm{a}})\\
&= \varepsilon_H - \tilde{Z}^{\mathrm{T}}\rho_2 + \mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2\mathbf{v}(\tilde{W}_{\mathrm{a}})
\end{aligned}$$

The above equations hold for both t and $\{t_j\}_{j=1}^{N}$. The time derivative of L_2 becomes

$$\begin{aligned}
\dot{L}_2 &= \frac{1}{N+1}\\
&\quad \times \left\{\frac{\tilde{Z}^{\mathrm{T}}\rho_2(t)}{m_2^2(t)}\left[\varepsilon_H(t) - \tilde{Z}^{\mathrm{T}}\rho_2(t) + \mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2(t)\mathbf{v}(\tilde{W}_{\mathrm{a}})\right]\right.\\
&\quad \left. + \sum_{j=1}^{N}\frac{\tilde{Z}^{\mathrm{T}}\rho_2(t_j)}{m_2^2(t_j)}\left[\varepsilon_H(t_j) - \tilde{Z}^{\mathrm{T}}\rho_2(t_j) + \mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2(t_j)\mathbf{v}(\tilde{W}_{\mathrm{a}})\right]\right\}
\end{aligned}$$

Note that since ϕ_{a} is bounded, we have

$$\mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}(R \otimes \phi_{\mathrm{a}}\phi_{\mathrm{a}}^{\mathrm{T}})\mathbf{v}(\tilde{W}_{\mathrm{a}}) \le k_1\left\|\mathbf{v}(\tilde{W}_{\mathrm{a}})\right\|^2 \le k_1\left\|\tilde{Z}\right\|^2$$

where $k_1 < \infty$ is a positive constant. Hence

$$\begin{aligned}
\mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2\mathbf{v}(\tilde{W}_{\mathrm{a}}) &\le k_1\int_{t-T}^{t} e^{-\gamma(\tau-t)}\left\|\tilde{Z}\right\|^2 \mathrm{d}\tau\\
&\le \frac{k_1}{\gamma\beta_3}(e^{\gamma \mathrm{T}} - 1)\tilde{Z}^{\mathrm{T}}\bar{\rho}_2\bar{\rho}_2^{\mathrm{T}}\tilde{Z}
\end{aligned}$$

where the second inequality is from the PE condition. Under the same condition, we further have

$$\left|\tilde{Z}^{\mathrm{T}}\frac{\rho_2}{m_2^2}\right| \leq \sqrt{\beta_4}\left\|\tilde{Z}\right\|$$

and

$$\left|\tilde{Z}^{\mathrm{T}}\frac{\rho_2}{m_2^2}\mathbf{v}(\tilde{W}_{\mathrm{a}})^{\mathrm{T}}D_2\mathbf{v}(\tilde{W}_{\mathrm{a}})\right| \leq \frac{\sqrt{\beta_4}k_1}{\gamma\beta_3}(e^{\gamma \mathrm{T}}-1)\left\|\tilde{Z}\right\| \times \tilde{Z}^{\mathrm{T}}\bar{\rho}_2\bar{\rho}_2^{\mathrm{T}}\tilde{Z}$$

Define a large bound B such that for arbitrary $\left\|\tilde{Z}\right\| \leq B$, we can select a small integral interval T to ensure $(\sqrt{\beta_4}k_1/\gamma\beta_3)(e^{\gamma \mathrm{T}}-1)\left\|\tilde{Z}\right\| \leq \varepsilon_T$, where $\varepsilon_T < 1$ is a positive constant. Consequently

$$\dot{L}_2 \leq -\frac{1-\varepsilon_T}{N+1}\lambda_{\min}(\bar{H}_2)\left\|\tilde{Z}\right\|^2 + b_H\left\|\tilde{Z}\right\|$$

where

$$\bar{H}_2 = \bar{\rho}_2(t)\bar{\rho}_2^{\mathrm{T}}(t) + \sum_{j=1}^{N}\bar{\rho}_2(t_j)\bar{\rho}_2^{\mathrm{T}}(t_j)$$

As a result, $\tilde{Z}$ converge to the residual set

$$\left\{\tilde{Z} : \left\|\tilde{Z}\right\| \leq \frac{(N+1)b_H}{(1-\varepsilon_T)\lambda_{\min}(\bar{H}_2)}\right\} \qquad \square$$

Remark 2: In Section 2, we have mentioned at the proper selection of γ and Q_{e}, the error dynamics can be stabilised by the optimal tracking controller, indicating it is possible to initialise an admissible policy u_0 that makes the system stable. During the online learning process, the control input is unchanged with $u = u_0 + e_u$. If u_0 is a stabilising policy and e_u is designed small enough, the stability of the system is guaranteed.

Remark 3: The convergence rate is improved if we can increase the minimum eigenvalue of $H_2 = \sum_{j=1}^{N}\bar{\rho}_2(t_j)\bar{\rho}_2^{\mathrm{T}}(t_j)$. So in the learning process, new observations keep updating some old data in the experience set in order to increase $\lambda_{\min}(H_2)$. In addition, from the proof of Theorem 3, the larger $\lambda_{\min}(\bar{H}_2)$ is, the smaller estimation error $\tilde{Z}$ is bounded.

Remark 4: Another existing literature that investigates ER for the optimal control is presented by Modares *et al.* in [30]. The CT non-linear optimal regulation problem is addressed in their work. Except that, the major difference is that [30] combines ER with an on-policy scheme while our algorithm is based on the off-policy method. On-policy method is expected to use the data generated by the currently estimated policy. When applying the past experience that is produced from other old policies, just like the design in [30], the accuracy of the results may be perturbed. For our off-policy algorithm, the difference between the executed policy and the estimated policy is explicitly considered in the HJB equation (12). So the ER technique performs a more positive role in the off-policy algorithm.

Remark 5: In the proof, an appropriate initialisation is required for $\hat{Z}$ to make $\left\|\tilde{Z}\right\| \leq B$. Such requirement can be satisfied if a prior policy evaluation process is conducted to learn $\hat{Z}$ under the tuning (18). Once a suitable $\hat{Z}$ is obtained, the adaptive optimal learning by (23) is followed. For some self-stable cases, such initialisation is unnecessary, like the experiments given in Section 5.

5 Experiment study

5.1 Linear case

To verify the performance, we first consider a spring, mass and damper system from [23] with the linear dynamics

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -5x_1 - 0.5x_2u$$
$$y = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}$$

The tracking trajectories are two sinusoidal signals $x_{\mathrm{d}1} = 0.5\sin\left(\sqrt{5}t\right)$, $x_{\mathrm{d}2} = 0.5\sqrt{5}\cos\left(\sqrt{5}t\right)$, generated from reference dynamics

$$\dot{x}_{\mathrm{d}} = \begin{bmatrix}0 & 1\\ -5 & 0\end{bmatrix}x_{\mathrm{d}}$$
$$y_{\mathrm{d}} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}x_{\mathrm{d}}$$

with starting state $x_{\mathrm{d}} = \left[0, 0.5\sqrt{5}\right]^{\mathrm{T}}$. The cost index is defined with $Q_{\mathrm{e}} = 10I$, $R = 1$ and the discounted factor γ selects 0.1. From (3), it is straightforward to obtain the augmented dynamics

$$\dot{X} = \begin{bmatrix}0 & 1 & 0 & 0\\ -5 & -0.5 & 0 & -0.5\\ 0 & 0 & 0 & 1\\ 0 & 0 & -5 & 0\end{bmatrix}X + \begin{bmatrix}0\\ 1\\ 0\\ 0\end{bmatrix}u = AX + Bu$$

Note that these dynamics are not provided to the algorithm.

Since the system is linear and the cost index is defined in quadratic, it is reasonable to infer that the value function is in the quadratic form, $V(X) = X^{\mathrm{T}}PX$, and the policy is expressed by $u = -K^{\mathrm{T}}X = -R^{-1}B^{\mathrm{T}}PX$. After inserting into the HJB equation (7), we get the following discounted algebraic Riccati equation (ARE)

$$A^{\mathrm{T}}P + PT - \gamma P + PBR^{-1}B^{\mathrm{T}}P + Q = 0$$

In our algorithm, we choose the basis functions for critic and actor NNs as

$$\phi_{\mathrm{c}}(X) = [X_1^2, X_2^2, X_3^2, X_4^2, X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4]^{\mathrm{T}}$$
$$\phi_{\mathrm{a}}(X) = [X_1, X_2, X_3, X_4]^{\mathrm{T}}$$

The learning rate α is set to 3000 and the experience set size selects $N = 20$. The control effort u applied to the system consists of several sinusoids with different frequencies. After 100s, the control input is turned to the converged tracker. The final actor weights are

$$\hat{W}_{\mathrm{a}} = [-0.770, -2.892, 0.002, 0.419]^{\mathrm{T}}$$

Using the ARE toolbox in MATLAB, we can solve the above ARE and get the ideal optimal tracker feedback gain

$$K = [0.769, 2.891, 0.022, -0.404]^{\mathrm{T}}$$

The evolution of the augmented states is depicted in Fig. 1. It is observed that once the learned tracker is applied to the system at 100s, tracking errors $e_{\mathrm{d}1}$ and $e_{\mathrm{d}2}$ are rapidly stabilised close to zero although reference signals x_{d} remain varying. The convergence of $\hat{W}_{\mathrm{c}}$ and $\hat{W}_{\mathrm{a}}$ is given in Figs. 2 and 3. Fig. 4 gives the detailed tracking trajectories under the tracker. The controlled system performs satisfying tracking trajectories to the desired ones, verifying the effectiveness of the algorithm. Fig. 5 presents the curve of the minimum eigenvalue of H_2 for the experience set. The grey vertical lines indicate the moments when newly observed data replace some old ones in the data set. The raise of the curve corresponds to the benefit of ER in the improvement of the convergence rate.

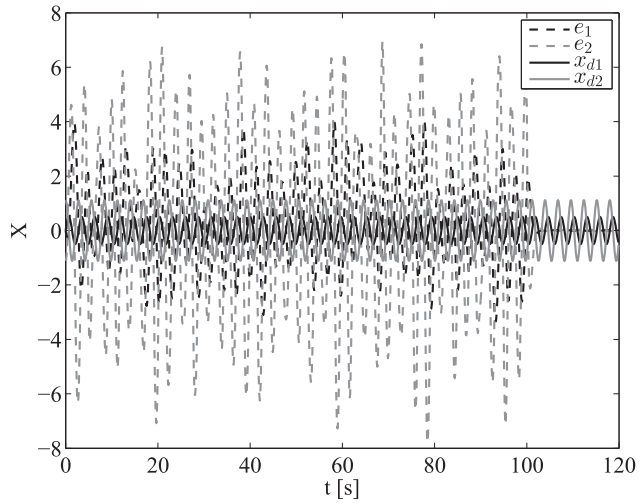

Fig. 1 *Evolution of the augmented system in the linear experiment*

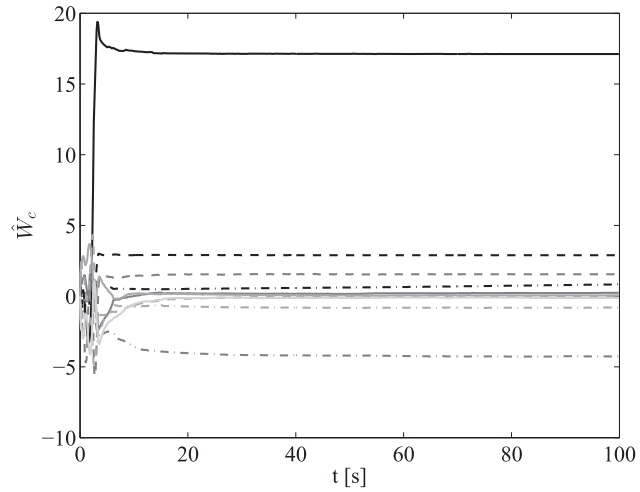

Fig. 2 *Convergence of the critic NN in the linear experiment*

5.2 Non-linear case

The second experiment is a non-linear system with dynamics

$$\begin{aligned}\dot{x}_1 &= -\sin(x_1) + x_2\\ \dot{x}_2 &= -x_1^3 + u\\ y &= x_1\end{aligned}$$

The desired trajectory for x_1 is given by $x_{\mathrm{d1}} = 0.1\cos(0.3t)$ from

$$\begin{aligned}\dot{x}_{\mathrm{d}} &= \begin{bmatrix} 0 & 0.3 \\ -0.3 & 0 \end{bmatrix} x_{\mathrm{d}}\\ y_{\mathrm{d}} &= x_{\mathrm{d1}}\end{aligned}$$

with the initial state $x_{\mathrm{d}} = [0.1, 0]^{\mathrm{T}}$. In this experiment, we select $Q_{\mathrm{e}} = 2.5$ and $R = 1$, while the other parameters are set the same as the linear case.

The evolution of the augmented system is plotted in Fig. 6. After 100s of the adaptive learning, the converged tracker replaces the exploratory control input, and the tracking error e_1 is regulated around zero. Figs. 7 and 8 depict the convergence

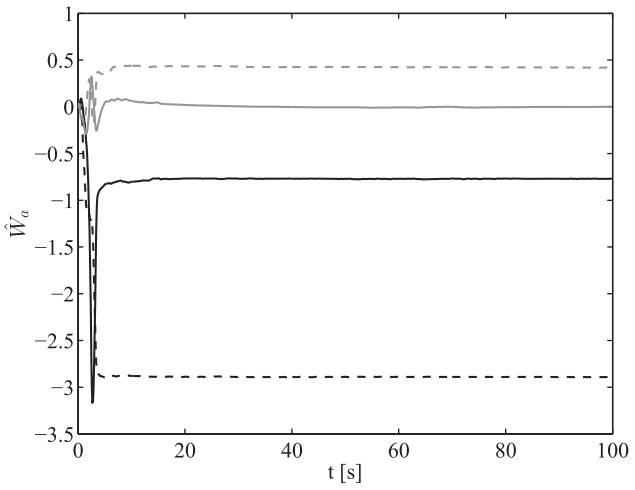

Fig. 3 *Convergence of the actor NN in the linear experiment*

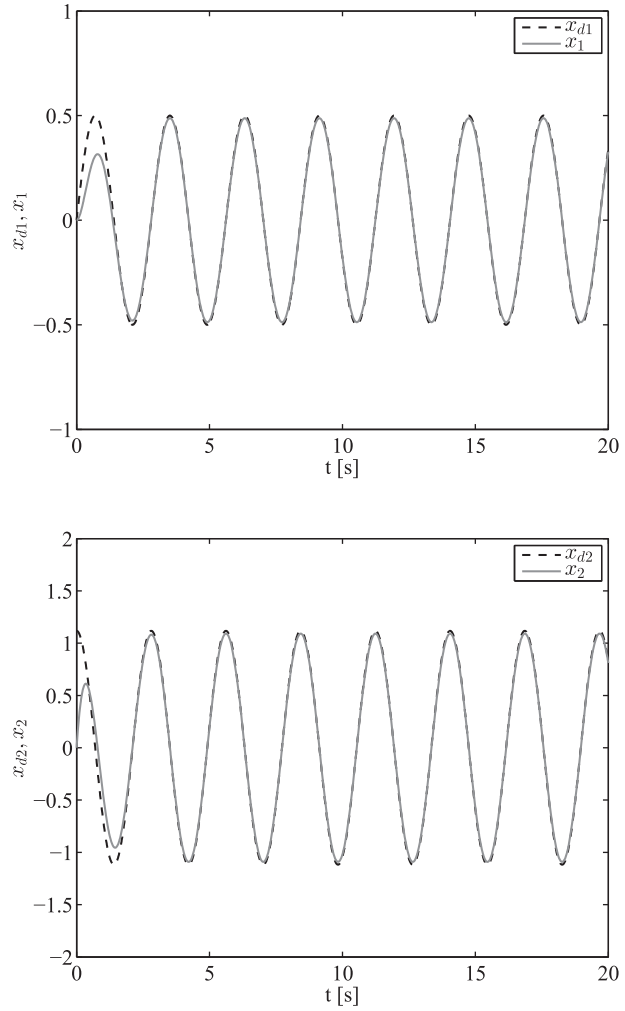

Fig. 4 *Tracking trajectories of the linear system under the learned tracker*

of $\hat{W}_{\mathrm{c}}$ and $\hat{W}_{\mathrm{a}}$, respectively. The final learned NN weights are $\hat{W}_{\mathrm{c}} = [1.064,\ 0.962,\ 1.187,\ 0.529,\ 1.023,\ -1.061, 0.669, -1.665, 0.835, -0.808]^{\mathrm{T}}$, $\hat{W}_{\mathrm{a}} = [-0.513, -0.962, 0.831, -0.414]^{\mathrm{T}}$. Fig. 9 presents the performance of the learned tracker. It is observed that x_1 closely tracks the cosine signal. The curve of the minimum eigenvalue of H_2 is given in Fig. 10.

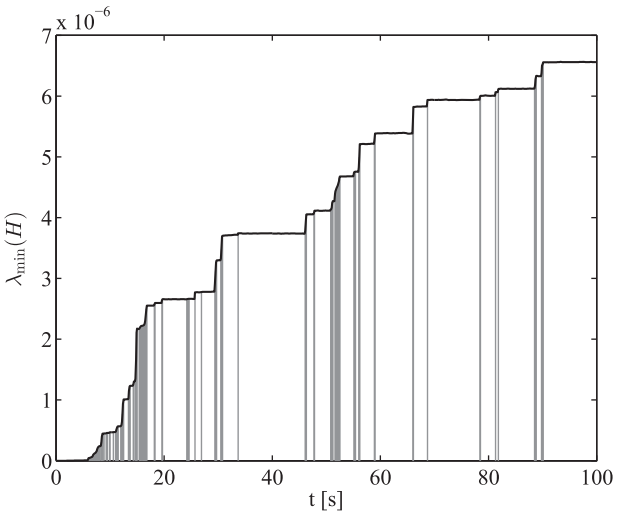

Fig. 5 *Curve of the minimum eigenvalue of H_2 in the linear experiment*

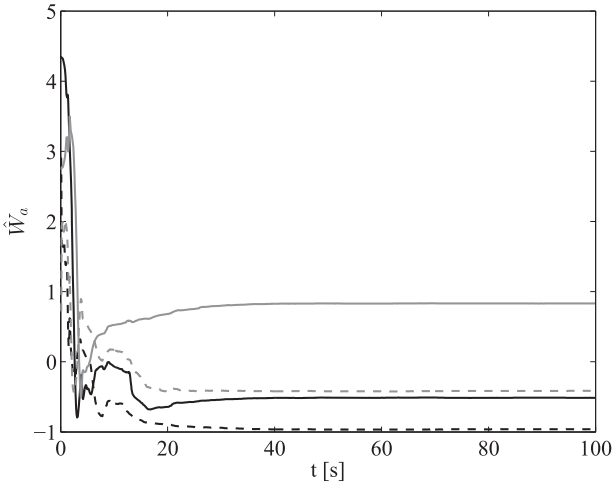

Fig. 8 *Convergence of the actor NN in the non-linear experiment*

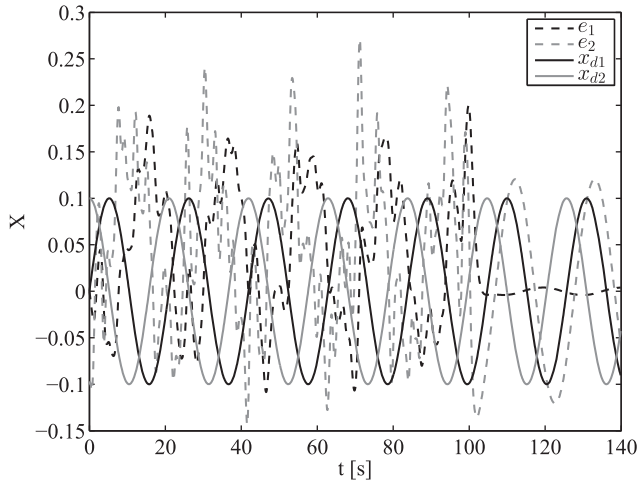

Fig. 6 *Evolution of the augmented system in the non-linear experiment*

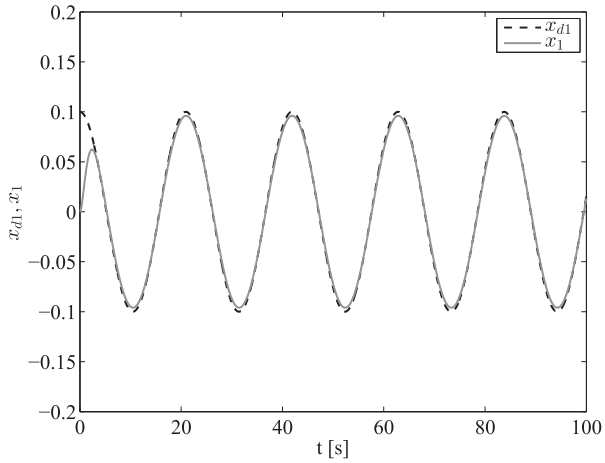

Fig. 9 *Tracking trajectories of the non-linear system under the learned tracker*

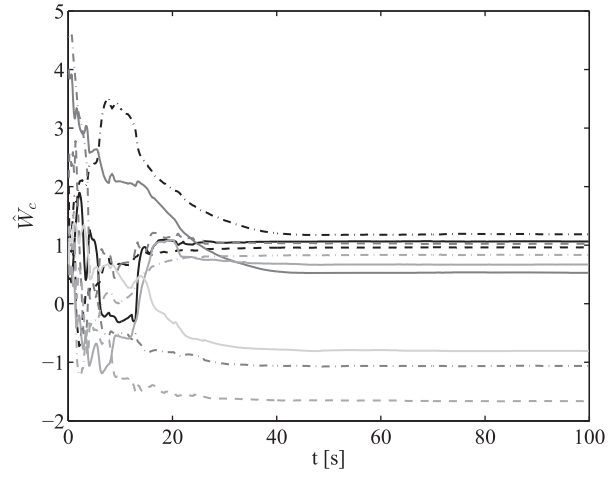

Fig. 7 *Convergence of the critic NN in the non-linear experiment*

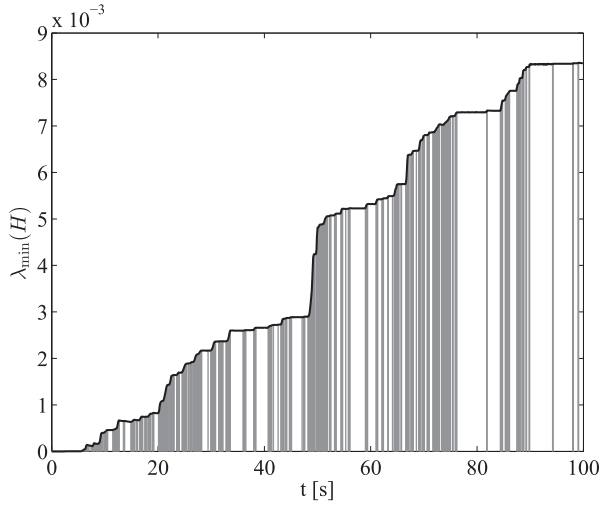

Fig. 10 *Curve of the minimum eigenvalue of H_2 in the non-linear experiment*

6 Conclusion

In this paper, we employ several RL techniques, including PI, IRL, actor–critic and ER, to form an online adaptive algorithm to solve the optimal tracking problem for CT non-linear systems with completely unknown dynamics. PI provides a feasible approach to the complex HJB equation; IRL eliminates the dependence on the system dynamics; actor–critic structure efficiently approximates the value and policy functions; and ER significantly improves the convergence rate. By constructing an augmented system, the tracking problem is converted to a regulation problem, which contributes to the application of the above methods.

7 Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under grants nos. 61273136, 61573353 and 61533017, by Beiing Nova Program under grant no. Z141101001814094, and by Science and Technology Foundation of SGCC under grant no. DG71-14-032.

8 References

1 Sutton, R.S., Barto, A.G.: 'Reinforcement learning: an introduction' (MIT Press, Cambridge, MA, 1998)
2 Lewis, F.L., Liu, D.: 'Reinforcement learning and approximate dynamic programming for feedback control' (Wiley, New York, 2012)
3 Powell, W.B.: 'Approximate dynamic programming: solving the curses of dimensionality' (Wiley-Interscience, 2007)
4 Zhang, H., Liu, D., Luo, Y., *et al.*: 'Adaptive dynamic programming for control. Algorithms and stability' (Springer-Verlag, London, 2012)
5 Murray, J.J., Cox, C.J., Lendaris, G.G., *et al.*: 'Adaptive dynamic programming', *IEEE Tran. Syst. Man Cybern. C, Appl. Rev.*, 2002, **32**, (2), pp. 140–153
6 Abu-Khalaf, M., Lewis, F.L.: 'Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach', *Automatica*, 2005, **41**, (5), pp. 779–791
7 Li, H., Liu, D.: 'Optimal control for discrete-time affine non-linear systems using general value iteration', *IET Control Theory Appl.*, 2012, **6**, (18), pp. 2725–2736
8 Yang, X., Liu, D., Wei, Q.: 'Online approximate optimal control for affine non-linear systems with unknown internal dynamics using adaptive dynamic programming', *IET Control Theory Appl.*, 2014, **8**, (16), pp. 1676–1688
9 Zhao, D., Zhu, Y.: 'MEC–a near-optimal online reinforcement learning algorithm for continuous deterministic systems', *IEEE Trans. Neural Netw. Learn. Syst.*, 2015, **26**, (2), pp. 346–356
10 Zhu, Y., Zhao, D., Liu, D.: 'Convergence analysis and application of fuzzy-HDP for nonlinear discrete-time HJB systems', *Neurocomputing*, 2015, **149**, Part A, pp. 124–131
11 Lewis, F., Vrabie, D.: 'Reinforcement learning and adaptive dynamic programming for feedback control', *IEEE Circuits Syst. Mag.*, 2009, **9**, (3), pp. 32–50
12 Wang, F.-Y., Zhang, H., Liu, D.: 'Adaptive dynamic programming: An introduction', *IEEE Comput. Intell. Mag.*, 2009, **4**, (2), pp. 39–47
13 Park, Y.-M., Choi, M.-S., Lee, K.: 'An optimal tracking neuro-controller for nonlinear dynamic systems', *IEEE Trans. Neural Netw.*, 1996, **7**, (5), pp. 1099–1110
14 Toussaint, G., Basar, T., Bullo, F.: 'H∞-optimal tracking control techniques for nonlinear underactuated systems', Proc. 39th IEEE Conf. on Decision and Control, 2000, vol.**3**, pp. 2078–2083
15 Alameda-Hernandez, E., Blanco, D., Ruiz, D., *et al.*: 'Optimal tracking of time-varying systems with the overdetermined recursive instrumental variable algorithm', *IET Control Theory Appl.*, 2007, **1**, (1), pp. 291–297
16 Jiang, X.-W., Guan, Z.-H., Feng, G., *et al.*: 'Optimal tracking performance of networked control systems with channel input power constraint', *IET Control Theory Appl.*, 2012, **6**, (11), pp. 1690–1698
17 Zhang, H., Wei, Q., Luo, Y.: 'A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm', *IEEE Trans. Syst. Man Cybern. B, Cybern.*, 2008, **38**, (4), pp. 937–942
18 Dierks, T., Jagannathan, S.: 'Optimal tracking control of affine nonlinear discrete-time systems with unknown internal dynamics'. Proc. 48th IEEE Conf. on Decision and Control, held jointly with the 28th Chinese Control Conf., December 2009SEP, pp. 6750–6755
19 Dierks, T., Jagannathan, S.: 'Optimal control of affine nonlinear continuous-time systems', 2010 American Control Conf. (ACC), June 2010, pp. 1568–1573
20 Zhang, H., Cui, L., Zhang, X., *et al.*: 'Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method', *IEEE Trans. Neural Netw.*, 2011, **22**, (12), pp. 2226–2236
21 Wang, D., Liu, D., Wei, Q.: 'Finite-horizon neuro-optimal tracking control for a class of discrete-time nonlinear systems using adaptive dynamic programming approach', *Neurocomputing*, 2012, **78**, (1), pp. 14–22
22 Kamalapurkar, R., Dinh, H., Bhasin, S., *et al.*: 'Approximate optimal trajectory tracking for continuous-time nonlinear systems', *Automatica*, 2015, **51**, pp. 40–48
23 Modares, H., Lewis, F.L.: 'Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning', *Automatica*, 2014, **50**, (7), pp. 1780–1792
24 Kiumarsi, B., Lewis, F.L., Modares, H., *et al.*: 'Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics', *Automatica*, 2014, **50**, (4), pp. 1167–1175
25 Vrabie, D., Lewis, F.: 'Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems', *Neural Netw.*, 2009, **22**, (3), pp. 237–246
26 Vamvoudakis, K.G., Vrabie, D., Lewis, F.L.: 'Online adaptive algorithm for optimal control with integral reinforcement learning', *Int. J. Robust Nonlinear Control*, 2014, **24**, (17), pp. 2686–2710
27 Jiang, Y., Jiang, Z.-P.: 'Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics', *Automatica*, 2012, **48**, (10), pp. 2699–2704
28 Lee, J., Park, J., Choi, Y.: 'Approximate dynamic programming for continuous-time linear quadratic regulator problems: relaxation of known input-coupling matrix assumption', *IET Control Theory Appl.*, 2012, **6**, (13), pp. 2063–2075
29 Luo, B., Wu, H.-N., Huang, T., Liu, D.: 'Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design', *Automatica*, 2014, **50**, (12), pp. 3281–3290
30 Modares, H., Lewis, F.L., Naghibi-Sistani, M.-B.: 'Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems', *Automatica*, 2014, **50**, (1), pp. 193–202
31 Modares, H., Lewis, F.: 'Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning', *IEEE Trans. Autom. Control*, 2014, **59**, (11), pp. 3051–3056
32 Li, C., Liu, D., Li, H.: 'Finite horizon optimal tracking control of partially unknown linear continuous-time systems using policy iteration', *IET Control Theory Appl.*, 2015, **9**, (12), pp. 1791–1801
33 Qin, C., Zhang, H., Luo, Y.: 'Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming', *Int. J. Control*, 2014, **87**, (5), pp. 1000–1009
34 Modares, H., Lewis, F., Jiang, Z.-P.: 'H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning', *IEEE Trans. Neural Netw. Learn. Syst.*, 2015, **26**, (10), pp. 2550–2562
35 Lee, J.Y., Park, J.B., Choi, Y.H.: 'Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations', *IEEE Trans. Neural Netw. Learn. Syst.*, 2015, **26**, (5), pp. 916–932