Neural Networks 71 (2015) 150–158

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Reinforcement learning solution for HJB equation arising in constrained optimal control problem

Biao Luo [a], Huai-Ning Wu [b], Tingwen Huang [c], Derong Liu [d,*]

[a] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] Science and Technology on Aircraft Control Laboratory, Beihang University (Beijing University of Aeronautics and Astronautics), Beijing 100191, China
[c] Texas A&M University at Qatar, PO Box 23874, Doha, Qatar
[d] School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

ARTICLE INFO

Article history:
Received 4 January 2015
Received in revised form 15 August 2015
Accepted 16 August 2015
Available online 24 August 2015

Keywords:
Constrained optimal control
Data-based
Off-policy reinforcement learning
Hamilton–Jacobi–Bellman equation
The method of weighted residuals

ABSTRACT

The constrained optimal control problem depends on the solution of the complicated Hamilton–Jacobi–Bellman equation (HJBE). In this paper, a data-based off-policy reinforcement learning (RL) method is proposed, which learns the solution of the HJBE and the optimal control policy from real system data. One important feature of the off-policy RL is that its policy evaluation can be realized with data generated by other behavior policies, not necessarily the target policy, which solves the insufficient exploration problem. The convergence of the off-policy RL is proved by demonstrating its equivalence to the successive approximation approach. Its implementation procedure is based on the actor–critic neural networks structure, where the function approximation is conducted with linearly independent basis functions. Subsequently, the convergence of the implementation procedure with function approximation is also proved. Finally, its effectiveness is verified through computer simulations.

1. Introduction

Optimal control is an important part of control theory, which has been widely investigated over the past several decades. The bottleneck of its applications to nonlinear systems is that it depends on the solution of the Hamilton–Jacobi–Bellman equation (HJBE) (Bertsekas, 2005; Hull, 2003; Lewis, Vrabie, & Syrmos, 2013), which is extremely difficult to obtain analytically. Over the past few years, reinforcement learning (RL) (Lendaris, 2009; Powell, 2007; Precup, Sutton, & Dasgupta, 2001; Sutton & Barto, 1998), has appeared as an efficient tool to solve the HJBE and many meaningful results (Faust, Ruymgaart, Salman, Fierro, & Tapia, 2014; Jiang & Jiang, 2012; Lee, Park, & Choi, 2012; Liu, Wang, & Li, 2014; Liu & Wei, 2014; Luo, Wu, Huang, & Liu, 2014; Modares & Lewis, 2014; Murray, Cox, Lendaris, & Saeks, 2002; Vamvoudakis & Lewis, 2010; Vrabie & Lewis, 2009; Vrabie, Pastravanu, Abu-Khalaf, & Lewis, 2009; Wang, Liu, & Li, 2014; Wei & Liu, 2012; Yang, Liu, & Wang, 2014; Yang, Liu, Wang, & Wei, 2014; Zhao, Xu, & Jagannathan, 2014) have been reported. For example, appropriate estimators were employed for approximating value function such that the temporal difference error is minimized (Doya, 2000). Murray et al. (2002) suggested two policy iteration algorithms that avoid the necessity of knowing the internal system dynamics. Vrabie et al. (2009) extended their result and proposed a new policy iteration algorithm to solve the linear quadratic regulation problem online along a single state trajectory. A nonlinear version of this algorithm was presented in Vrabie and Lewis (2009) by using neural network (NN) approximator. Vamvoudakis and Lewis (2010) also gave a so-called synchronous policy iteration algorithm which tunes synchronously the weight parameters of both NNs in the actor–critic structure. An integral reinforcement learning (IRL) method (Modares & Lewis, 2014) was introduced to solve the linear quadratic tracking problem of partially-unknown continuous-time systems. Online adaptive optimal control (Jiang & Jiang, 2012) and Q-learning (Lee et al., 2012) algorithms were developed for linear quadratic regulator problem. Off-policy RL approaches were proposed to solve the nonlinear data-based optimal control problem (Luo et al., 2014) and partially model-free H_∞ control problem (Luo, Wu, & Huang, 2015). However, it is noted that control constraints are not involved in these results.

In practice, constraints widely exist in real control systems and have damaging effects on the system performance, and thus should

* Corresponding author.
E-mail addresses: biao.luo@hotmail.com (B. Luo), whn@buaa.edu.cn (H.-N. Wu), tingwen.huang@qatar.tamu.edu (T. Huang), derong@ustb.edu.cn (D. Liu).

http://dx.doi.org/10.1016/j.neunet.2015.08.007
0893-6080/

be accounted for during the controller design process. For the constrained optimal control problem, several results (Abu-Khalaf & Lewis, 2005; He & Jagannathan, 2005, 2007; Heydari & Balakrishnan, 2013; Liu, Wang, & Yang, 2013; Lyshevski, 1998; Modares, Lewis, & Naghibi-Sistani, 2013; Zhang, Luo, & Liu, 2009) have been reported recently. A nonquadratic cost functional was introduced by Lyshevski (1998) to confront input constraints, and then the associated HJBE was reformulated accordingly. As the extensions of the method in Saridis and Lee (1979) and Beard, Saridis, and Wen (1997) to handle constrained optimal control problem, model-based successive approximation method was used for solving the HJBE of continuous-time systems (Abu-Khalaf & Lewis, 2005) and discrete-time systems (Chen & Jagannathan, 2008). Modares, Lewis, and Naghibi-Sistani (2014) developed an experience-replay based IRL algorithm for nonlinear partially unknown constrained-input systems. A heuristic dynamic programming was used to solve the constrained optimal control problem of nonlinear discrete-time systems (Zhang et al., 2009). The single network based adaptive critics method was proposed for finite-horizon nonlinear constrained optimal control design (Heydari & Balakrishnan, 2013). However, the data-based constrained nonlinear optimal control problem is rarely studied with off-policy RL and still remains an open issue.

In this paper, a data-based off-policy RL method is proposed for learning the constrained optimal control policy form real system data instead of using mathematical model. The rest of the paper is arranged as follows. Section 2 gives the problem description and Section 3 presents a model-based successive approximation method. The data-based off-policy RL method is developed in Section 4. Section 5 shows the simulation results and Section 6 gives the conclusions.

Notation: $\mathbb{R}$ and $\mathbb{R}^n$ are the set of real numbers and the n-dimensional Euclidean space, respectively. $\|\cdot\|$ denotes the vector norm or matrix norm in $\mathbb{R}^n$. The superscript T is used for the transpose and I denotes the identify matrix of appropriate dimension. $\bigtriangledown \triangleq \partial/\partial x$ denotes a gradient operator. $C^1(\mathcal{X})$ is a function space on $\mathcal{X}$ with continuous first derivatives. Let $\mathcal{X}$ and $\mathcal{U}$ be compact sets, denote $\mathcal{D} \triangleq \{(x,u)|x \in \mathcal{X}, u \in \mathcal{U}\}$. For column vector functions $s_1(x,u)$ and $s_2(x,u)$, where $(x,u) \in \mathcal{D}$, define inner product $\langle s_1(x,u), s_2(x,u)\rangle_{\mathcal{D}} \triangleq \int_{\mathcal{D}} s_1^{\mathrm{T}}(x,u)s_2(x,u)\mathrm{d}(x,u)$ and norm $\|s_1(x,u)\|_{\mathcal{D}} \triangleq \langle s_1(x,u), s_2(x,u)\rangle_{\mathcal{D}}^{1/2}$.

2. Problem description

Let us consider the following continuous-time nonlinear system:

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t), \quad x(0) = x_0, \tag{1}$$

where $x = [x_1, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^n$ is the state, x_0 is the initial state and $u = [u_1, \ldots, u_m]^{\mathrm{T}} \in \mathbb{R}^m$ is the control input constrained by $|u_i| \leqslant \beta$, $\beta > 0$. Assume that $f(x) + g(x)u(x)$ is Lipschitz continuous on $\mathcal{X}$ that contains the origin, $f(0) = 0$, and the system is stabilizable on $\mathcal{X}$, i.e., there exists a continuous control function $u(x)$ such that the system is asymptotically stable. $f(x)$ and $g(x)$ are continuous *unknown* vector or matrix functions of appropriate dimensions.

The optimal control problem under consideration is to learn a state feedback control law $u(t) = u(x(t))$ from real system data, such that the system (1) is closed-loop asymptotically stable, and minimize the following generalized infinite horizon cost functional:

$$V(x_0) \triangleq \int_0^{+\infty} (Q(x(t)) + W(u(t)))\mathrm{d}t, \tag{2}$$

where $Q(x)$ and $W(u)$ are positive definite functions, i.e., for $\forall x \neq 0$, $u \neq 0$, $Q(x) > 0$, $W(u) > 0$, and $Q(x) = 0$, $W(u) = 0$ only when $x = 0$, $u = 0$. Then, the optimal control problem is briefly presented as

$$u(t) = u^*(x) \triangleq \arg\min_u V(x_0). \tag{3}$$

3. Model-based successive approximation method

For the model-based optimal control problem (3), i.e., the mathematical models of $f(x)$ and $g(x)$ are completely known, it can be converted to solving the HJBE. In Abu-Khalaf and Lewis (2005), a model-based successive approximation method was given for solving the HJBE, where the HJBE is successively approximated by a sequence of linear partial differential equations. Before we start, the definition of admissible control (Abu-Khalaf & Lewis, 2005; Beard et al., 1997) is given.

Definition 1 (*Admissible Control*)**.** For the given system (1), $x \in \mathcal{X}$, a control policy $u(x)$ is defined to be admissible with respect to the cost function (2) on $\mathcal{X}$, denoted by $u(x) \in \mathfrak{U}(\mathcal{X})$, if, (1) u is continuous on $\mathcal{X}$, (2) $u(0) = 0$, (3) $u(x)$ stabilizes the system, and (4) $V(x) < \infty, \forall x \in \mathcal{X}$. □

For $\forall u(x) \in \mathfrak{U}(\mathcal{X})$, its value function $V(x)$ of (2) satisfies the following linear partial differential equation (Abu-Khalaf & Lewis, 2005):

$$[\nabla V(x)]^{\mathrm{T}}(f(x) + g(x)u(x)) + Q(x) + W(u) = 0, \tag{4}$$

where $V(x) \in C^1(\mathcal{X})$, $V(x) \geq 0$ and $V(0) = 0$. From the optimal control theory (Anderson & Moore, 2007; Bertsekas, 2005; Lewis et al., 2013), if using the optimal control $u^*(x)$, the Eq. (4) results in the HJBE

$$[\nabla V^*(x)]^{\mathrm{T}}(f(x) + g(x)u^*(x)) + Q(x) + W(u^*) = 0. \tag{5}$$

For the system (1) with input constraints $|u_i| \leqslant \beta$, the following nonquadratic form $W(u)$ for the cost functional (2) can be used (Abu-Khalaf & Lewis, 2005; Lyashevskiy, 1996; Lyshevski, 1998; Modares et al., 2013):

$$W(u) = 2\sum_{l=1}^{m} r_l \int_0^{u_l} \varphi^{-1}(\mu_l)\mathrm{d}\mu_l, \tag{6}$$

where $\mu \in \mathbb{R}^m$, $r_l > 0$ and $\varphi(\cdot)$ is a continuous one-to-one bounded function satisfying $|\varphi(\cdot)| \leqslant \beta$ with $\varphi(0) = 0$. Moreover, $\varphi(\cdot)$ is a monotonic odd function and its derivative is bounded. An example of $\varphi(\cdot)$ is the hyperbolic tangent $\tanh(\cdot)$. Denoting $R = \mathrm{diag}(r_1, \ldots, r_m)$, it follows from Abu-Khalaf and Lewis (2005) and Lyshevski (1998) that the HJBE (5) of the constrained optimal control problem is given by

$$[\nabla V^*]^{\mathrm{T}}\left(f - g\varphi\left(\frac{1}{2}R^{-1}g^{\mathrm{T}}\nabla V^*\right)\right) + Q(x) + W\left(-\varphi\left(\frac{1}{2}R^{-1}g^{\mathrm{T}}\nabla V^*\right)\right) = 0. \tag{7}$$

By solving the HJBE for $V^*(x)$, the optimal control policy is obtained as:

$$u^*(x) = -\varphi\left(\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla V^*(x)\right). \tag{8}$$

For simplicity of description, define

$$\nu^*(x) \triangleq -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla V^*(x). \tag{9}$$

Then, the HJBE (7) and optimal control (8) can briefly be rewritten as:

$$(\nabla V^*)^{\mathrm{T}}\left(f+g\varphi(v^*)\right)+Q+W(\varphi(v^*))=0 \tag{10}$$

$$u^*=\varphi(v^*). \tag{11}$$

In Abu-Khalaf and Lewis (2005), the HJBE (10) is successively approximated with a sequence of iterative equations

$$[\nabla V^{(i)}]^{\mathrm{T}}(f+gu^{(i)})+Q+W(u^{(i)})=0;\quad i=0,1,\ldots, \tag{12}$$

with policy improvement

$$u^{(i+1)}=\varphi(v^{(i+1)}), \tag{13}$$

where

$$v^{(i+1)}\triangleq-\frac{1}{2}R^{-1}g^{\mathrm{T}}\nabla V^{(i)}. \tag{14}$$

Remark 1. From the famous RL books (Bertsekas, 2005; Sutton & Barto, 1998), policy iteration is a basic framework of RL. Over the past few years, policy iteration has already been employed to solve the unconstrained optimal control problems of linear systems (Jiang & Jiang, 2012; Modares & Lewis, 2014; Vrabie et al., 2009) and nonlinear systems (Vamvoudakis & Lewis, 2010; Vrabie & Lewis, 2009). In fact, the successive approximation between the iterative equations (12) and (13) is essentially a model-based policy iteration method, which involves two basic steps: policy evaluation and policy improvement. The Eq. (12) is policy evaluation for evaluating the control policy $u^{(i)}$ for its value function $V^{(i)}$, and Eq. (13) is policy improvement for obtaining an improved control policy $u^{(i+1)}$ based on the current value function $V^{(i)}$. By implementing the two steps alternatively, it has been proven in Abu-Khalaf and Lewis (2005) that the value function $V^{(i)}$ will converge to the solution of the HJBE (10), i.e., $\lim_{i\to\infty}V^{(i)}=V^*$ and thus $\lim_{i\to\infty}u^{(i)}=u^*$. Note that the iterative equation (12) involves the full mathematical system models of $f(x)$ and $g(x)$. In Abu-Khalaf and Lewis (2005), a model-based approach was developed to solve the iterative equation (12) by using NN for approximating the value function $V^{(i)}$. □

4. Data-based constrained optimal control

For the data-based constrained optimal control problem under consideration, the explicit expression of the HJBE (10) is unavailable since the mathematical system models $f(x)$ and $g(x)$ are unknown, which prevents using model-based approaches for control design. To overcome this difficulty, a data-based off-policy RL is developed to learn the optimal control policy.

4.1. Off-policy reinforcement learning

In this subsection, the off-policy RL approach is derived based on (12) and (13). Inspired by Jiang and Jiang (2014) and Luo et al. (2014), the system (1) can be rewritten as

$$\dot{x}=f+gu^{(i)}+g[u-u^{(i)}] \tag{15}$$

for $\forall u\in\mathcal{U}$. Let us consider the case when $V^{(i)}(x)$ be the solution of the iterative equation (12). By using (12)–(14), we take derivative of $V^{(i)}(x)$ with respect to time along the state of system (15)

$$\begin{aligned}\frac{\mathrm{d}V^{(i)}(x)}{\mathrm{d}t}&=[\nabla V^{(i)}]^{\mathrm{T}}(f+gu^{(i)})-[\nabla V^{(i)}]^{\mathrm{T}}g[u^{(i)}-u]\\&=-Q-W(u^{(i)})+2(v^{(i+1)})^{\mathrm{T}}R[u^{(i)}-u]\\&=-Q-W(\varphi(v^{(i)}))+2(v^{(i+1)})^{\mathrm{T}}R[\varphi(v^{(i)})-u].\end{aligned} \tag{16}$$

Integrating both sides of (16) on the interval $[t,t+\Delta t]$ and rearranging terms yields,

$$\begin{aligned}&2\int_t^{t+\Delta t}[v^{(i+1)}(x(\tau))]^{\mathrm{T}}R[\varphi(v^{(i)}(x(\tau)))-u(\tau)]\mathrm{d}\tau\\&\quad+V^{(i)}(x(t))-V^{(i)}(x(t+\Delta t))\\&=\int_t^{t+\Delta t}\left(Q(x(\tau))+W(\varphi(v^{(i)}(x(\tau))))\right)\mathrm{d}\tau,\end{aligned} \tag{17}$$

where $V^{(i)}(x)$ is an unknown function and $v^{(i+1)}(x)$ is an unknown function vector to be solved. The main idea of the off-policy RL is solving the iterative equation (17) instead of the iterative equation (12). Compared with the iterative equation (12), the iterative equation (17) does not require the explicit mathematical model of the system (1), i.e., $f(x)$ and $g(x)$. According to the definition of off-policy RL (Maei, Szepesvári, Bhatnagar, & Sutton, 2010; Precup et al., 2001; Sutton & Barto, 1998), the value function $V^{(i)}$ of the target control policy $u^{(i)}$ can be evaluated by using system data generated with other behavior policies u and not restricted to the target policy. This implies that the proposed off-policy RL method has an advantage that it can learn the value function and control policy from system data that are generated according to more exploratory or even random policies.

For the proposed off-policy RL, its aim is to learn the constrained optimal control policy (8) by iteratively solving Eq. (17) for the unknown function $V^{(i)}(x)$ and function vector $v^{(i+1)}(x)$. Thus, it is necessary to prove that the generated sequences $\{V^{(i)}\}$ and $\{v^{(i)}\}$ will converge to V^* and v^*, respectively.

Theorem 1. *Let $V^{(i)}(x)\in C^1(\mathcal{X})$, $V^{(i)}(x)\geq 0$, $V^{(i)}(0)=0$ and $\varphi(v^{(i+1)}(x))\in\mathfrak{U}(\mathcal{X})$. $(V^{(i)}(x),v^{(i+1)}(x))$ is the solution of Eq. (17) if and only if it is the solution of the iterative equations (12)–(14), i.e., Eq. (17) is equivalent to the iterative equations (12)–(14).*

Proof. From the derivation of Eq. (17), we have that if $(V^{(i)},v^{(i+1)})$ is the solution of the iterative equations (12)–(14), it also satisfies Eq. (17). To complete the proof, we have to show that $(V^{(i)},v^{(i+1)})$ is the unique solution of Eq. (17). The proof is by contradiction.

Before start, we derive a simple fact. Consider

$$\begin{aligned}&\lim_{\Delta t\to0}\frac{1}{\Delta t}\int_t^{t+\Delta t}\hbar(\tau)\mathrm{d}\tau\\&=\lim_{\Delta t\to0}\frac{1}{\Delta t}\left(\int_0^{t+\Delta t}\hbar(\tau)\mathrm{d}\tau-\int_0^t\hbar(\tau)\mathrm{d}\tau\right)\\&=\frac{\mathrm{d}}{\mathrm{d}t}\int_0^t\hbar(\tau)\mathrm{d}\tau\\&=\hbar(t).\end{aligned} \tag{18}$$

From (17), we have

$$\begin{aligned}&\frac{\mathrm{d}V^{(i)}(x)}{\mathrm{d}t}\\&=\lim_{\Delta t\to0}\frac{1}{\Delta t}\left(V^{(i)}(x(t+\Delta t))-V^{(i)}(x(t))\right)\\&=2\lim_{\Delta t\to0}\int_t^{t+\Delta t}[v^{(i+1)}(x(\tau))]^{\mathrm{T}}\\&\quad\times R[\varphi(v^{(i)}(x(\tau)))-u(\tau)]\mathrm{d}\tau\\&\quad-\lim_{\Delta t\to0}\frac{1}{\Delta t}\int_t^{t+\Delta t}\left(Q(x(\tau))+W(\varphi(v^{(i)}(x(\tau))))\right)\mathrm{d}\tau.\end{aligned} \tag{19}$$

By using the fact (18), Eq. (19) is rewritten as

$$\begin{aligned}\frac{\mathrm{d}V^{(i)}(x)}{\mathrm{d}t}&=2[v^{(i+1)}(x(t))]^{\mathrm{T}}R[\varphi(v^{(i)}(x(t)))-u(t)]\\&\quad-Q(x(t))-W(\varphi(v^{(i)}(x(t)))).\end{aligned} \tag{20}$$

Suppose that $(W(x), \upsilon(x))$ is another solution of Eq. (17), where $W(x) \in C^1(\mathcal{X})$ with boundary condition $W(0) = 0$ and $\varphi(\upsilon(x)) \in \mathfrak{U}(\mathcal{X})$. Thus, (W, υ) also satisfies Eq. (20), i.e.,

$$\frac{\mathrm{d}W(x)}{\mathrm{d}t} = 2\upsilon^{\mathrm{T}}(x(t))R[\varphi(\nu^{(i)}(x(t))) - u(t)] - Q(x(t)) - W(\varphi(\nu^{(i)}(x(t)))). \tag{21}$$

Substituting Eq. (21) from (20) yields,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(V^{(i)}(x) - W(x)\right) = 2[\nu^{(i+1)}(x(t)) - \upsilon(x(t))]^{\mathrm{T}} \times R[\varphi(\nu^{(i)}(x(t))) - u(t)]. \tag{22}$$

This means that Eq. (22) holds for $\forall u \in \mathcal{U}$. If letting $u = \varphi(\nu^{(i)})$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(V^{(i)}(x) - W(x)\right) = 0. \tag{23}$$

This implies that $V^{(i)}(x) - W(x) = c$ holds for $\forall x \in \mathcal{X}$, where c is a real constant. For $x = 0$, $c = V^{(i)}(0) - W(0) = 0$. Then, $V^{(i)}(x) - W(x) = 0$, i.e., $W(x) = V^{(i)}(x)$ for $\forall i, x \in \mathcal{X}$. From (22), we have

$$[\nu^{(i+1)}(x) - \upsilon(x)]^{\mathrm{T}}R[\varphi(\nu^{(i)}(x)) - u] = 0$$

for $\forall u \in \mathcal{U}$. Thus, $\nu^{(i+1)}(x) - \upsilon(x) = 0$, i.e., $\upsilon(x) = \nu^{(i+1)}(x)$ for $\forall i$, $x \in \mathcal{X}$. This completes the proof. □

Theorem 1 shows that the off-policy RL with iterative equation (17) is theoretically equivalent to the model-based successive approximation method with the iterative equations (12)–(14), which is convergent as proved in Abu-Khalaf and Lewis (2005). Thus, the convergence of the off-policy RL can be guaranteed.

4.2. The method of weighted residuals

At each step of the off-policy RL, it requires to solve the iterative equation (17) for $V^{(i)}(x)$ and $\nu^{(i+1)}(x) = [\nu_1^{(i+1)}(x), \ldots, \nu_m^{(i+1)}(x)]^{\mathrm{T}}$. By using real system data, the method of weighted residuals (MWR) is derived based on the actor–critic NN structure. Although similar MWR can also be found in Luo et al. (2014), for the sake of clearness and completeness, the MWR will be developed for (17) which is much more complicated and different to some extent. Let $\Psi(x) \triangleq \{\psi_j(x)\}_{j=1}^{\infty}$ and $\Phi^l(x) \triangleq \{\phi_k^l(x)\}_{k=1}^{\infty}$ be complete sets of any linearly independent basis functions, such that $\psi_j(0) = 0$ and $\phi_k^l(0) = 0$ for $\forall l, j, k$. Then, the solution $(V^{(i)}(x), \nu^{(i+1)}(x))$ of the iterative equation (17) can be expressed as linear combination of basis function sets $\Psi(x)$ and $\Phi^l(x)$, i.e., $V^{(i)}(x) = \sum_{j=1}^{\infty}\theta_{V,j}^{(i)}\psi_j(x)$ and $\nu_l^{(i+1)}(x) = \sum_{k=1}^{\infty}\theta_{\nu_l,k}^{(i+1)}\phi_k^l(x)$, $(l = 1, \ldots, m)$, which are assumed to converge pointwise in $\mathcal{X}$. By using finite-dimensional sets $\Psi_N(x) \triangleq [\psi_1(x), \ldots, \psi_{L_V}(x)]^{\mathrm{T}}$ and $\Phi_N^l(x) \triangleq [\phi_1^l(x), \ldots, \phi_{L_u}^l(x)]^{\mathrm{T}}$ as neuron activation functions, the real output of critic and actor NNs can be, respectively, given by

$$\hat{V}^{(i)}(x) = \Psi_N^{\mathrm{T}}(x)\hat{\theta}_V^{(i)} \tag{24}$$

$$\hat{\nu}_l^{(i+1)}(x) = (\Phi_N^l(x))^{\mathrm{T}}\hat{\theta}_{\nu_l}^{(i+1)}, \tag{25}$$

where $\hat{\theta}_V^{(i)} \triangleq [\hat{\theta}_{V,1}^{(i)}, \ldots, \hat{\theta}_{V,L_V}^{(i)}]^{\mathrm{T}}$ and $\hat{\theta}_{\nu_l}^{(i+1)} \triangleq [\hat{\theta}_{\nu_l,1}^{(i+1)}, \ldots, \hat{\theta}_{\nu_l,L_u}^{(i+1)}]^{\mathrm{T}}$ are the estimations of the unknown ideal weight vectors $\theta_V^{(i)} \triangleq [\theta_{V,1}^{(i)}, \ldots, \theta_{V,L_V}^{(i)}]^{\mathrm{T}}$ and $\theta_{\nu_l}^{(i+1)} \triangleq [\theta_{\nu_l,1}^{(i+1)}, \ldots, \theta_{\nu_l,L_u}^{(i+1)}]^{\mathrm{T}}$. The expression (25) can be rewritten as a compact form

$$\hat{\nu}^{(i+1)}(x) = \left[\hat{\nu}_1^{(i+1)}(x), \ldots, \hat{\nu}_m^{(i+1)}(x)\right]^{\mathrm{T}} = \left[(\Phi_N^1(x))^{\mathrm{T}}\hat{\theta}_{u_1}^{(i+1)}, \ldots, (\Phi_N^m(x))^{\mathrm{T}}\hat{\theta}_{u_m}^{(i+1)}\right]^{\mathrm{T}}. \tag{26}$$

Due to the truncation error of the trail solutions (24) and (25), the replacement of $V^{(i)}$ and $\nu^{(i+1)}$ in the iterative equation (17) with $\hat{V}^{(i)}$ and $\hat{\nu}^{(i+1)}$ respectively, yields the following residual error:

$$\begin{aligned}\sigma^{(i)}(x(t), u(t)) \triangleq\ & 2\int_t^{t+\Delta t}[\hat{\nu}^{(i+1)}(x(\tau))]^{\mathrm{T}} \times R[\varphi(\hat{\nu}^{(i)}(x(\tau))) - u(\tau)]\mathrm{d}\tau \\ & + \hat{V}^{(i)}(x(t)) - \hat{V}^{(i)}(x(t+\Delta t)) \\ & - \int_t^{t+\Delta t}\left(Q(x(\tau)) + W(\varphi(\hat{\nu}^{(i)}(x(t))))\right)\mathrm{d}\tau.\end{aligned} \tag{27}$$

By using (24) and (26), we have

$$\begin{aligned}&\sigma^{(i)}(x(t), u(t)) \\ &= [\Psi_N(x(t)) - \Psi_N(x(t+\Delta t))]^{\mathrm{T}}\hat{\theta}_V^{(i)} \\ &\quad + 2\sum_{l=1}^{m} r_l\left[\int_t^{t+\Delta t}\varphi\left(\left(\Phi_N^l(x(t))\right)^{\mathrm{T}}\hat{\theta}_{\nu_l}^{(i)}\right) \times \left(\Phi_N^l(x(\tau))\right)^{\mathrm{T}}\mathrm{d}\tau\right]\hat{\theta}_{\nu_l}^{(i+1)} \\ &\quad - 2\sum_{l=1}^{m} r_l\left[\int_t^{t+\Delta t}u_l(\tau)\left(\Phi_N^l(x(\tau))\right)^{\mathrm{T}}\mathrm{d}\tau\right]\hat{\theta}_{\nu_l}^{(i+1)} \\ &\quad - \int_t^{t+\Delta t}Q(x(\tau))\mathrm{d}\tau \\ &\quad - 2\sum_{l=1}^{m} r_l\int_t^{t+\Delta t}\left(\int_0^{\varphi\left(\left(\Phi_N^l(x(t))\right)^{\mathrm{T}}\hat{\theta}_{\nu_l}^{(i)}\right)}\varphi^{-1}(\mu_l)\mathrm{d}\mu_l\right)\mathrm{d}\tau.\end{aligned} \tag{28}$$

For simplicity of notation, define

$$\begin{aligned}\rho_{\Delta\Psi}(x(t)) &\triangleq [\Psi_N(x(t)) - \Psi_N(x(t+\Delta t))]^{\mathrm{T}} \\ \rho_{\Phi}^{(i)l}(x(t)) &\triangleq \int_t^{t+\Delta t}\varphi\left(\left(\Phi_N^l(x(t))\right)^{\mathrm{T}}\hat{\theta}_{\nu_l}^{(i)}\right)\left(\Phi_N^l(x(\tau))\right)^{\mathrm{T}}\mathrm{d}\tau \\ \rho_{u\Phi}^{l}(x(t), u(t)) &\triangleq \int_t^{t+\Delta t}u_l(\tau)\left(\Phi_N^l(x(\tau))\right)^{\mathrm{T}}\mathrm{d}\tau \\ \rho_{Q}(x(t)) &\triangleq \int_t^{t+\Delta t}Q(x(\tau))\mathrm{d}\tau \\ \rho_{1}^{(i)l}(x(t)) &\triangleq \int_t^{t+\Delta t}\left(\int_0^{\varphi\left(\left(\Psi_N^l(x(t))\right)^{\mathrm{T}}\hat{\theta}_{\nu_l}^{(i)}\right)}\varphi^{-1}(\mu_l)\mathrm{d}\mu_l\right)\mathrm{d}\tau.\end{aligned} \tag{29}$$

Then, Eq. (28) is rewritten as

$$\begin{aligned}\sigma^{(i)}(x(t), u(t)) =\ & \rho_{\Delta\Psi}(x(t))\hat{\theta}_V^{(i)} + 2\sum_{l=1}^{m} r_l\rho_{\Phi}^{(i)l}(x(t))\hat{\theta}_{\nu_l}^{(i+1)} \\ & - 2\sum_{l=1}^{m} r_l\rho_{u\Phi}^{l}(x(t), u(t))\hat{\theta}_{\nu_l}^{(i+1)} \\ & - \rho_Q(x(t)) - 2\sum_{l=1}^{m} r_l\rho_1^{(i),l}(x(t)).\end{aligned} \tag{30}$$

To write Eq. (30) in a compact form, define

$$\begin{aligned}\hat{\theta}^{(i+1)} &\triangleq \left[\left(\hat{\theta}_V^{(i)}\right)^{\mathrm{T}}, \left(\hat{\theta}_{u_1}^{(i+1)}\right)^{\mathrm{T}}, \ldots, \left(\hat{\theta}_{u_m}^{(i+1)}\right)^{\mathrm{T}}\right] \\ \overline{\rho}_{u\Phi}^{(i)l}(x(t), u(t)) &\triangleq r_l\left[\rho_{\Phi}^{(i)l}(x(t)) - \rho_{u\Phi}^{l}(x(t), u(t))\right] \\ \overline{\rho}^{(i)}(x(t), u(t)) &\triangleq \left[\rho_{\Delta\Psi}^{\mathrm{T}}(x(t)), 2\overline{\rho}_{u\Phi}^{(i)1}(x(t), u(t)), \ldots, 2\overline{\rho}_{u\Phi}^{(i)m}(x(t), u(t))\right] \\ \pi^{(i)}(x(t)) &\triangleq \rho_Q(x(t)) + 2\sum_{l=1}^{m} r_l\rho_1^{(i)l}(x(t))\end{aligned} \tag{31}$$

where $\hat{\theta}^{(i+1)}$ is the unknown constant vector of size $L = L_V + mL_u$. Then, the Eq. (30) is rewritten as

$$\sigma^{(i)}(x(t), u(t)) = \overline{\rho}^{(i)}(x(t), u(t))\hat{\theta}^{(i+1)} - \pi^{(i)}(x(t)), \tag{32}$$

where we denote $\overline{\rho}^{(i)} = [\overline{\rho}_1^{(i)}, \ldots, \overline{\rho}_L^{(i)}]^{\mathrm{T}}$. In the MWR, the unknown constant vector $\hat{\theta}^{(i+1)}$ can be solved in such a way that the residual error $\sigma^{(i)}(x, u)$ (for $\forall t \geqslant 0$) of (32) is forced to be zero in some average sense. The weighted integrals of the residual are set to zero:

$$\left\langle \mathcal{W}_L^{(i)}(x, u), \sigma^{(i)}(x, u) \right\rangle_{\mathcal{D}} = 0, \tag{33}$$

where $\mathcal{W}_L^{(i)}(x, u) \triangleq [\omega_1^{(i)}(x, u), \ldots, \omega_L^{(i)}(x, u)]^{\mathrm{T}}$ is named the weighted function vector. Then, the substitution of (32) into (33) yields,

$$\left\langle \mathcal{W}_L^{(i)}(x, u), \overline{\rho}^{(i)}(x, u) \right\rangle_{\mathcal{D}} \hat{\theta}^{(i+1)} - \left\langle \mathcal{W}_L^{(i)}(x, u), \pi^{(i)}(x) \right\rangle_{\mathcal{D}} = 0,$$

where the notations $\left\langle \mathcal{W}_L^{(i)}, \overline{\rho}^{(i)} \right\rangle_{\mathcal{D}}$ and $\left\langle \mathcal{W}_L^{(i)}, \pi^{(i)} \right\rangle_{\mathcal{D}}$ are given by

$$\left\langle \mathcal{W}_L^{(i)}, \overline{\rho}^{(i)} \right\rangle_{\mathcal{D}} \triangleq \begin{bmatrix} \left\langle \omega_1^{(i)}, \overline{\rho}_1^{(i)} \right\rangle_{\mathcal{D}} & \cdots & \left\langle \omega_1^{(i)}, \overline{\rho}_L^{(i)} \right\rangle_{\mathcal{D}} \\ \vdots & \ddots & \vdots \\ \left\langle \omega_L^{(i)}, \overline{\rho}_1^{(i)} \right\rangle_{\mathcal{D}} & \cdots & \left\langle \omega_L^{(i)}, \overline{\rho}_L^{(i)} \right\rangle_{\mathcal{D}} \end{bmatrix}$$

and

$$\left\langle \mathcal{W}_L^{(i)}, \pi^{(i)} \right\rangle_{\mathcal{D}} \triangleq \left[\left\langle \omega_1^{(i)}, \pi^{(i)} \right\rangle_{\mathcal{D}}, \ldots, \left\langle \omega_L^{(i)}, \pi^{(i)} \right\rangle_{\mathcal{D}} \right]^{\mathrm{T}}.$$

Thus, $\hat{\theta}^{(i+1)}$ can be obtained with

$$\hat{\theta}^{(i+1)} = \left\langle \mathcal{W}_L^{(i)}, \overline{\rho}^{(i)} \right\rangle_{\mathcal{D}}^{-1} \left\langle \mathcal{W}_L^{(i)}, \pi^{(i)} \right\rangle_{\mathcal{D}}. \tag{34}$$

Note that the computations of $\left\langle \mathcal{W}_L^{(i)}, \overline{\rho}^{(i)} \right\rangle_{\mathcal{D}}$ and $\left\langle \mathcal{W}_L^{(i)}, \pi^{(i)} \right\rangle_{\mathcal{D}}$ involve many numerical integrals on domain $\mathcal{D}$, which are computationally expensive. Thus, the Monte-Carlo integration method (Luo et al., 2014; Peter Lepage, 1978) is introduced, which is especially competitive on multi-dimensional domain. We now illustrate the Monte-Carlo integration for computing $\left\langle \mathcal{W}_L^{(i)}(x, u), \overline{\rho}^{(i)}(x, u) \right\rangle_{\mathcal{D}}$. Let $I_{\mathcal{D}} \triangleq \int_{\mathcal{D}} \mathrm{d}(x, u)$, and $\mathcal{S}_M \triangleq \{(x_k, u_k) | (x_k, u_k) \in \mathcal{D}, k = 1, 2, \ldots, M\}$ be the set that are sampled on domain $\mathcal{D}$, where M is the size of sample set $\mathcal{S}_M$. Generally, it is desired to collect data set $\mathcal{S}_M$ as rich as possible to cover the domain $\mathcal{D}$ as much as possible. With the sample set $\mathcal{S}_M$, $\left\langle \mathcal{W}_L^{(i)}(x, u), \overline{\rho}^{(i)}(x, u) \right\rangle_{\mathcal{D}}$ is approximately computed with

$$\begin{aligned} \left\langle \mathcal{W}_L^{(i)}(x, u), \overline{\rho}^{(i)}(x, u) \right\rangle_{\mathcal{D}} &= \int_{\mathcal{D}} \mathcal{W}_L^{(i)}(x, u)\overline{\rho}^{(i)}(x, u)\mathrm{d}(x, u) \\ &= \frac{I_{\mathcal{D}}}{M} \sum_{k=1}^{M} \mathcal{W}_L^{(i)}(x_k, u_k)\overline{\rho}^{(i)}(x_k, u_k) \\ &= \frac{I_{\mathcal{D}}}{M} (W^{(i)})^{\mathrm{T}} Z^{(i)}, \end{aligned} \tag{35}$$

where $W^{(i)} \triangleq \left[\mathcal{W}_L^{(i)}(x_1, u_1), \ldots, \mathcal{W}_L^{(i)}(x_M, u_M) \right]^{\mathrm{T}}$ and $Z^{(i)} \triangleq \left[\overline{\rho}^{(i)\mathrm{T}}(x_1, u_1), \ldots, \overline{\rho}^{(i)\mathrm{T}}(x_M, u_M) \right]^{\mathrm{T}}$. Similarly,

$$\begin{aligned} \left\langle \mathcal{W}_L^{(i)}(x, u), \pi^{(i)}(x) \right\rangle_{\mathcal{D}} &= \frac{I_{\mathcal{D}}}{M} \sum_{k=1}^{M} \left(\mathcal{W}_L^{(i)}(x_k, u_k) \right)^{\mathrm{T}} \pi^{(i)}(x_k) \\ &= \frac{I_{\mathcal{D}}}{M} (W^{(i)})^{\mathrm{T}} \eta^{(i)}, \end{aligned} \tag{36}$$

where $\eta^{(i)} \triangleq \left[\pi^{(i)}(x_1), \ldots, \pi^{(i)}(x_M) \right]^{\mathrm{T}}$. Then, the substitution of (35) and (36) into (34) yields,

$$\hat{\theta}^{(i+1)} = \left[(W^{(i)})^{\mathrm{T}} Z^{(i)} \right]^{-1} (W^{(i)})^{\mathrm{T}} \eta^{(i)}. \tag{37}$$

It is observed that the sample set $\mathcal{S}_M$ is generated on domain $\mathcal{D}$, based on which $W^{(i)}$, $Z^{(i)}$ and $\eta^{(i)}$ can be computed and then the unknown parameter vector $\hat{\theta}^{(i+1)}$ is obtained with the expression (37) accordingly. With $\hat{\theta}^{(i+1)}$, the unknown function $V^{(i)}(x)$ and function vector $\nu^{(i+1)}(x)$ can be approximately obtained by expressions (24) and (26), respectively.

Remark 2. It is assumed that the system state x is measurable and sufficient system data can be collected for implementing the off-policy RL. For the function approximation of $V^{(i)}(x)$ and $\nu^{(i+1)}(x)$, linearly independent basis functions are required, which may bring limitation to some extent. Further studies will be conducted to reduce the deficiencies of this limitation. □

4.3. Implementation of off-policy reinforcement learning

Based on the parameter update strategy (37), we give the implementation procedure of the off-policy RL for data-based constrained optimal control design.

Algorithm 1. Off-policy RL for data-based constrained optimal control design.

- *Step 1*: Collect real system data (x_k, u_k) for sample set $\mathcal{S}_M$, and then compute $\rho_{\Delta\Psi}(x_k)$, $\rho_{u\Phi}^l(x_k, u_k)$ and $\rho_Q(x_k)$;
- *Step 2*: Give initial parameter vectors $\hat{\theta}_u^{(0)}$ such that $\phi(\hat{\nu}^{(0)}) \in \mathfrak{U}(\mathcal{X})$. Let $i = 0$;
- *Step 3*: Compute $W^{(i)}$, $Z^{(i)}$ and $\eta^{(i)}$, and update $\hat{\theta}^{(i+1)}$ with (37);
- *Step 4*: Let $i = i + 1$. If $\|\hat{\theta}^{(i)} - \hat{\theta}^{(i-1)}\| \leq \xi$ (ξ is a small positive number), stop iteration and $\hat{\theta}_u^{(i)}$ is employed to obtain the final control policy $\phi(\hat{\nu}^{(i)})$; else go back to Step 3 and continue. □

Remark 3. Note that the off-policy RL method (i.e., Algorithm 1) is also suitable for solving the unconstrained optimal control problem. By using a sufficiently large control bound β, the constrained optimal control problem is relaxed to be an unconstrained one. □

Remark 4. On-policy RL is one of the popular methods used for control design (Modares & Lewis, 2014; Vrabie & Lewis, 2009; Vrabie et al., 2009). To evaluate the value function of a general target control policy μ in the on-policy RL methods, it needs to generate system data using the policy μ. This biases the learning process by under-representing states that are unlikely to occur under μ. As a result, the estimated value function of these underrepresented states may be highly inaccurate, and seriously impact the improved policy. This is known as inadequate exploration—a particularly acute difficult issue in RL methods, which is rarely discussed in the existing works using RL techniques for control design. On the other hand, for real implementation of the on-policy learning methods, the approximate target control policy $\hat{\mu}$ (rather than the actual target policy μ) is usually used to generate data for learning its value function. In other words, the on-policy learning methods use the "inaccurate" data to learn its value function, which will increase the accumulated error. Note that these mentioned problems are solved in the developed off-policy RL method (i.e., Algorithm 1). The policy evaluation in the off-policy RL method can be realized with data generated by other behavior policies while not necessarily the target policy, which increases the "exploration" ability during the learning process. Moreover, in the off-policy RL algorithm, the control u and state x can be arbitrary on $\mathcal{U}$ and $\mathcal{X}$, where no error occurs during the process of generating data, and thus the accumulated error can be reduced. □

4.4. Convergence analysis with approximation

Under NN approximation structure (24) and (25), the convergence of the developed off-policy RL method is proved in the following Theorem 2.

Theorem 2. *Let the parameter vector $\hat{\theta}_k$ be computed with (37). Assume that there exist constants $\overline{M}_1 > 0$ and $\delta_1 > 0$, such that for $\forall M \geqslant \overline{M}_1$ and $i \geqslant 0$,*

$$\frac{1}{M}(W^{(i)})^{\mathrm{T}} Z^{(i)} \geqslant \delta_1 I_L. \tag{38}$$

For $\forall x \in \mathcal{X}$, $\varepsilon > 0$, there exist integers $L_1, L_2, I_1 > 0$ such that if $L_V \geqslant L_1, L_u \geqslant L_2$ and $i \geqslant I_1$, then

(1) $|\hat{V}^{(i)}(x) - V^{(i)}(x)| \leqslant \varepsilon$ *and* $\|\hat{\nu}^{(i)}(x) - \nu^{(i)}(x)\| \leqslant \varepsilon$.
(2) $|\hat{V}^{(i)}(x) - V^{*}(x)| \leqslant \varepsilon$ *and* $\|\hat{\nu}^{(i)}(x) - \nu^{*}(x)\| \leqslant \varepsilon$.

Proof. (1) Let $\overline{V}^{(i)}(x)$ be the solution of the following equation

$$[\nabla \overline{V}^{(i)}]^{\mathrm{T}}(f + g\hat{u}^{(i)}) + Q + W(\hat{u}^{(i)}) = 0, \tag{39}$$

where $\overline{V}^{(i)}(0) = 0$ and $\hat{u}^{(i)} = \varphi(\hat{\nu}^{(i)})$. Define

$$\overline{\nu}^{(i+1)} \triangleq -\frac{1}{2} R^{-1} g^{\mathrm{T}} \nabla \overline{V}^{(i)}. \tag{40}$$

From Section 4.2, $\overline{V}^{(i)}(x)$ and $\overline{\nu}^{(i+1)}(x)$ can be represented as $\overline{V}^{(i)}(x) = \sum_{j=1}^{\infty} \overline{\theta}_{V,j}^{(i)} \psi_j(x)$ and $\overline{\nu}_l^{(i+1)}(x) = \sum_{k=1}^{\infty} \overline{\theta}_{\nu_l,k}^{(i+1)} \phi_k^l(x)$, respectively. Define the error weight vector $\widetilde{\theta}^{(i+1)}$ as

$$\widetilde{\theta}^{(i+1)} \triangleq \hat{\theta}^{(i+1)} - \overline{\theta}^{(i+1)}. \tag{41}$$

Then, it follows from (37) and (41) that

$$(W^{(i)})^{\mathrm{T}} Z^{(i)} \hat{\theta}^{(i+1)} = (W^{(i)})^{\mathrm{T}} \eta^{(i)},$$

i.e.,

$$(W^{(i)})^{\mathrm{T}} Z^{(i)} \widetilde{\theta}^{(i+1)} = (W^{(i)})^{\mathrm{T}} \eta^{(i)} - (W^{(i)})^{\mathrm{T}} Z^{(i)} \overline{\theta}^{(i+1)}. \tag{42}$$

Multiplying $[\widetilde{\theta}^{(i+1)}]^{\mathrm{T}}$ on both sides of (42) yields

$$[\widetilde{\theta}^{(i+1)}]^{\mathrm{T}} (W^{(i)})^{\mathrm{T}} Z^{(i)} \widetilde{\theta}^{(i+1)} = [W^{(i)} \widetilde{\theta}^{(i+1)}]^{\mathrm{T}} [\eta^{(i)} - Z^{(i)} \overline{\theta}^{(i+1)}]. \tag{43}$$

By using (38), the left side of (43) satisfies

$$[\widetilde{\theta}^{(i+1)}]^{\mathrm{T}} (W^{(i)})^{\mathrm{T}} Z^{(i)} \widetilde{\theta}^{(i+1)} \geqslant M\delta_1 \|\widetilde{\theta}^{(i+1)}\|^2. \tag{44}$$

Combining (43) and (44) yields

$$M\delta_1 \|\widetilde{\theta}^{(i+1)}\|^2 \leqslant [W^{(i)} \widetilde{\theta}^{(i+1)}]^{\mathrm{T}} [\eta^{(i)} - Z^{(i)} \overline{\theta}^{(i+1)}]. \tag{45}$$

Based on definitions of $\eta^{(i)}$, $Z^{(i)}$ and $\theta^{(i+1)}$, the right side of (45) is given by

$$\begin{aligned}
&[W^{(i)} \widetilde{\theta}^{(i+1)}]^{\mathrm{T}} [\eta^{(i)} - Z^{(i)} \overline{\theta}^{(i+1)}] \\
&= [\widetilde{\theta}^{(i+1)}]^{\mathrm{T}} \sum_{l=1}^{M} \mathcal{W}_L(x_l, u_l) \pi^{(i)}(x_l) \\
&\quad - [\widetilde{\theta}^{(i+1)}]^{\mathrm{T}} \sum_{l=1}^{M} \mathcal{W}_L(x_l, u_l) \left[\overline{\rho}^{(i)}(x_l, u_l)\right]^{\mathrm{T}} \overline{\theta}^{(i+1)} \\
&= [\widetilde{\theta}^{(i+1)}]^{\mathrm{T}} \sum_{l=1}^{M} \mathcal{W}_L(x_l, u_l) \left[2 \int_t^{t+\Delta t} [\overline{\nu}^{(i+1)}(x_l(\tau))]^{\mathrm{T}} \right. \\
&\quad \times R[\varphi(\hat{\nu}^{(i)}(x_l(\tau))) - u_l(\tau)] \mathrm{d}\tau \\
&\quad + \overline{V}^{(i)}(x_l(t)) - \overline{V}^{(i)}(x_l(t + \Delta t)) \\
&\quad \left. - \int_t^{t+\Delta t} \left(Q(x_l(\tau)) + W(\varphi(\hat{\nu}^{(i)}(x_l(t))))\right) \mathrm{d}\tau \right] \\
&\quad + [\widetilde{\theta}^{(i+1)}]^{\mathrm{T}} \sum_{l=1}^{M} \mathcal{W}_L(x_l, u_l) e_l \\
&= [\widetilde{\theta}^{(i+1)}]^{\mathrm{T}} \sum_{l=1}^{M} \mathcal{W}_L(x_l, u_l) e_l,
\end{aligned} \tag{46}$$

where

$$\begin{aligned}
e_l = &- \sum_{k=L_V+1}^{\infty} \overline{\theta}_{V,k}^{(i)} [\psi_j(x_l(t)) - \psi_j(x_l(t + \Delta t))] \\
&- 2 \sum_{j=1}^{m} r_j \sum_{k=L_u+1}^{\infty} \overline{\theta}_{\nu_l,k}^{(i+1)} \int_t^{t+\Delta t} \phi_k^j(x_l(\tau)) \\
&\times [\varphi_j(\hat{\nu}^{(i)}(x_l(\tau))) - u_{j,l}(\tau)] \mathrm{d}\tau.
\end{aligned}$$

From (45) and (46), we have

$$\begin{aligned}
M\delta_1 \|\widetilde{\theta}^{(i+1)}\|^2 &\leqslant \left[W^{(i)} \widetilde{\theta}^{(i+1)}\right]^{\mathrm{T}} \left[\eta^{(i)} - Z^{(i)} \overline{\theta}^{(i+1)}\right] \\
&\leqslant \|\widetilde{\theta}^{(i+1)}\| \sum_{l=1}^{M} \|\mathcal{W}_L(x_l, u_l)\| \, |e_l| \\
&= M \|\widetilde{\theta}^{(i+1)}\| \varepsilon_1 \varepsilon_2
\end{aligned}$$

i.e.,

$$\|\widetilde{\theta}^{(i+1)}\| \leqslant \frac{1}{\delta_1} \varepsilon_1 \varepsilon_2, \tag{47}$$

where $\varepsilon_1 \triangleq \max |e_l|$ and $\varepsilon_2 \triangleq \max \|\mathcal{W}_L(x, u)\|$. Note that $\lim_{L_V, L_u \to \infty} e_l = 0$. Then, $\lim_{L_V, L_u \to \infty} \varepsilon_1 = 0$, i.e., $\lim_{L_V, L_u \to \infty} \|\widetilde{\theta}^{(i+1)}\| = 0$. Considering

$$\hat{V}^{(i)}(x) - \overline{V}^{(i)}(x) = \sum_{j=1}^{L_V} \widetilde{\theta}_{V,j}^{(i)} \psi_j(x) + \sum_{j=L_V+1}^{\infty} \overline{\theta}_{V,j}^{(i)} \psi_j(x),$$

$$\hat{\nu}_l^{(i+1)}(x) - \overline{\nu}_l^{(i+1)}(x) = \sum_{k=1}^{L_u} \widetilde{\theta}_{\nu_l,k}^{(i+1)} \phi_k^l(x) + \sum_{k=L_u+1}^{\infty} \overline{\theta}_{\nu_l,k}^{(i+1)} \phi_k^l(x),$$

we have

$$\lim_{L_V, L_u \to \infty} \hat{V}^{(i)}(x) = \overline{V}^{(i)}(x), \tag{48}$$

$$\lim_{L_V, L_u \to \infty} \hat{\nu}_l^{(i+1)}(x) = \overline{\nu}_l^{(i+1)}(x). \tag{49}$$

Next, we will use mathematical induction to prove that $\lim_{L_V, L_u \to \infty} \overline{V}^{(i)}(x) = V^{(i)}(x)$ and $\lim_{L_V, L_u \to \infty} \overline{\nu}_l^{(i+1)}(x) = \nu_l^{(i+1)}(x)$ for $i = 0, 1, 2 \ldots$.

(a) For $i = 0$, it follows from definitions of $\overline{V}^{(i)}(x)$ and $\overline{\nu}_l^{(i+1)}(x)$ that $\overline{V}^{(0)}(x) = V^{(0)}(x)$ and $\overline{\nu}_l^{(1)}(x) = \nu_l^{(1)}(x)$.

(b) For some i, assume that $\lim_{L_V, L_u \to \infty} \overline{V}^{(i-1)}(x) = V^{(i-1)}(x)$ and $\lim_{L_V, L_u \to \infty} \overline{\nu}_l^{(i)}(x) = \nu_l^{(i)}(x)$. According to (16) and (39),

$$\frac{\mathrm{d}\overline{V}^{(i)}(x)}{\mathrm{d}t} = -Q - W(\varphi(\hat{\nu}^{(i)})) + 2(\overline{\nu}^{(i+1)})^{\mathrm{T}} R[\varphi(\hat{\nu}^{(i)}) - u]. \tag{50}$$

From (16) and (50),

$$\begin{aligned}
&\overline{V}^{(i)}(x(t)) - V^{(i)}(x(t)) \\
&= \int_t^{\infty} [W(\varphi(\hat{\nu}^{(i)})) - W(\varphi(\nu^{(i)}))] \mathrm{d}\tau
\end{aligned}$$

$$+2\int_t^\infty (\overline{\nu}^{(i+1)})^{\mathrm{T}} R[\varphi(\hat{\nu}^{(i)}) - \varphi(\nu^{(i)})]\mathrm{d}\tau$$
$$+2\int_t^\infty [\overline{\nu}^{(i+1)} - \nu^{(i+1)}]^{\mathrm{T}} R[\varphi(\nu^{(i)}) - u]\mathrm{d}\tau. \quad (51)$$

According to (49) and $\lim_{L_V, L_u\to\infty} \overline{\nu}_l^{(i)}(x) = \nu_l^{(i)}(x)$, we have that $\lim_{L_V, L_u\to\infty} \hat{\nu}_l^{(i)}(x) = \nu_l^{(i)}(x)$. Then,

$$\lim_{L_V, L_u\to\infty} \overline{V}^{(i)}(x) = V^{(i)}(x), \quad (52)$$
$$\lim_{L_V, L_u\to\infty} \overline{\nu}_l^{(i+1)}(x) = \nu_l^{(i+1)}(x). \quad (53)$$

Based on Eqs. (48), (49), (52) and (53), for $\forall x \in \mathcal{X}$, $\varepsilon > 0$, there exist integers $L_1, L_2 > 0$ such that if $L_V \geqslant L_1$ and $L_u \geqslant L_2$. Then,

$$\begin{aligned} |\hat{V}^{(i)}(x) - V^{(i)}(x)| &\leqslant |\hat{V}^{(i)}(x) - \overline{V}^{(i)}(x)| + |\overline{V}^{(i)}(x) - V^{(i)}(x)| \\ &\leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \\ \|\hat{\nu}^{(i)}(x) - \nu^{(i)}(x)\| &\leqslant \|\hat{\nu}^{(i)}(x) - \overline{\nu}^{(i)}(x)\| + \|\overline{\nu}^{(i)}(x) - \nu^{(i)}(x)\| \\ &\leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

(2) From Ref. Abu-Khalaf and Lewis (2005) and Theorem 1, for $\forall \varepsilon > 0$, there exists integer I_1 such that for $\forall i \geqslant I_1$,

$$|V^{(i)}(x) - V^*(x)| \leqslant \frac{\varepsilon}{2}. \quad (54)$$

According to the part (1) of Theorem 1, there exist $L_V \geqslant L_1$ and $L_u \geqslant L_2$ such that

$$|\hat{V}^{(i)}(x) - V^{(i)}(x)| \leqslant \frac{\varepsilon}{2}. \quad (55)$$

From (54) and (55), we have

$$\begin{aligned} |\hat{V}^{(i)}(x) - V^*(x)| &\leqslant |\hat{V}^{(i)}(x) - V^{(i)}(x)| + |V^{(i)}(x) - V^*(x)| \\ &\leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Similarly, $\|\hat{\nu}^{(i)}(x) - \nu^*(x)\| \leqslant \varepsilon$. The proof is completed. □

5. Simulation studies

In this section, we study the effectiveness of the developed off-policy RL approach on a complex rotational/translational actuator (RTAC) nonlinear benchmark problem (Abu-Khalaf, Lewis, & Huang, 2008). The dynamics of the nonlinear plant poses challenges as the rotational and translational motions are coupled. In the simulation studies, select the weighted function vector as $w_L^{(i)}(x, u) = \overline{\rho}^{(i)}(x, u)$. Then, $W^{(i)} = Z^{(i)}$ and the parameter vector update strategy (37) becomes a least-square scheme. The RTAC system is given as follows:

$$\dot{x} = \begin{bmatrix} x_2 \\ \dfrac{-x_1 + \zeta x_4^2 \sin x_3}{1 - \zeta^2 \cos^2 x_3} \\ x_4 \\ \dfrac{\zeta \cos x_3 (x_1 - \zeta x_4^2 \sin x_3)}{1 - \zeta^2 \cos^2 x_3} \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{-\zeta \cos x_3}{1 - \zeta^2 \cos^2 x_3} \\ 0 \\ \dfrac{1}{1 - \zeta^2 \cos^2 x_3} \end{bmatrix} u,$$

$x_0 = [0.4, 0.0, 0.4, 0.0]^{\mathrm{T}}$,

where $\zeta = 0.2$. The input is constrained by $|u| \leqslant \beta$, where $\beta = 0.2$. Let $\varphi(\mu) = \beta \tanh(\mu/\beta)$ and $R = 1$, then $W(u)$ in cost functional (2) is

$$\begin{aligned} W(u) &= 2\int_0^u \beta \tanh^{-1}(\mu/\beta) R \mathrm{d}\mu \\ &= 2\beta R u \tanh^{-1}(u/\beta) + \beta^2 R \ln(1 - u^2/\beta^2). \end{aligned}$$

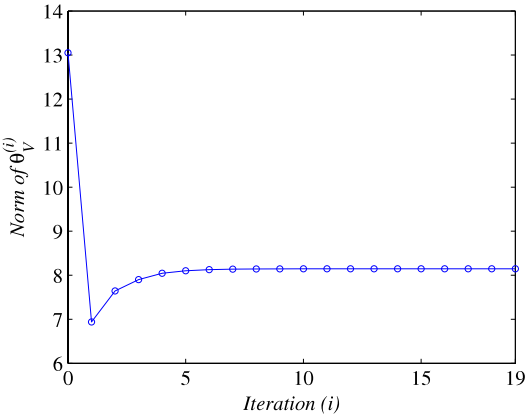

Fig. 1. The norm $\|\hat{\theta}_V^{(i)}\|$ at each iteration.

Let $Q(x) = x^{\mathrm{T}} S x$ with $S = \mathrm{diag}(0.5, 0.05, 0.05, 0.05)$. To learn the constrained optimal control policy with the off-policy RL method (Algorithm 1), select the basis function vectors as $\Psi_N(x) = [x_1^2, x_1x_2, x_1x_3, x_1x_4, x_2^2, x_2x_3, x_2x_4, x_3^2, x_3x_4x_1^2x_4^2, x_1x_2^3, x_1x_2^2x_3, x_1x_2^2x_4, x_1x_2x_3^2, x_1x_2x_3x_4, x_1x_2^2x_4, x_1x_3^3, x_1x_3^2x_4, x_1x_3x_4^2, x_1x_4^3, x_2^4, x_2^3x_3, x_2^2x_3^2, x_2x_3x_4, x_2^2x_4^2, x_2x_3^3, x_2x_3^2x_4, x_2x_4^3, x_3^4, x_3^3x_4, x_3^2x_4^2, x_3x_4^3, x_4^4]^{\mathrm{T}}$ with the size of $L_V = 42$, and $\Phi_N(x) = [x_1, x_2, x_3, x_4, \varphi^{\mathrm{T}}(x)]^{\mathrm{T}}$ with the size of $L_u = 46$, and initial $\hat{\theta}_u^{(0)}$ as $\hat{\theta}_u^{(0)} = [1.0, 1.0, -0.7, -2.0, 0, \ldots, 0]^{\mathrm{T}}$.

Collect sample set $\mathcal{S}_M$ with size $M = 1001$, and compute $\rho_{\Delta\Psi}(x_k)$, $\rho_Q(x_k)$, $\rho_{u\Phi}^l(x_k, u_k)$. Setting $\xi = 10^{-5}$, the simulation results show that at the 20th iteration (i.e., $i = 20$), the weight vectors converge, respectively, to $\hat{\theta}_V^{(20)} = [4.2970, 0.1216, -0.2196, -0.8742, 4.0075, 0.2672, 0.7472, 0.1643, 0.4953, 0.7819, 0.8625, 1.9686, 2.1268, 0.5542, -1.6487, -0.3671, -0.6932, -2.4891, -0.9257, 0.5616, 1.4230, 0.4928, 0.6333, 0.4141, 0.4928, 0.0537, 0.1653, 0.9117, 1.4092, 0.4641, -1.0433, -0.1432, 0.0280, 1.0835, 0.1701, -0.5234, -0.5237, -0.0486, -0.0376, 0.2384, -0.0521, -0.5238]^{\mathrm{T}}$ and $\hat{\theta}_u^{(20)} = [0.4421, 0.4591, -0.2291, -0.7333, 0.1905, -0.1791, -0.0575, -0.2978, -0.0232, -0.0739, 0.1672, -0.0035, -0.0009, 0.0519, 0.4666, -4.6554, -3.1500, -0.9666, 1.2378, 0.2776, 0.1788, 8.9946, 3.1199, 3.2885, 2.5886, -3.9747, 3.9289, -18.3913, -3.9747, 4.6712, -5.3350, -6.1230, 14.1708, -1.4422, 0.7945, 3.5809, 0.7768, 2.1957, 1.9014, -1.5518, -7.0940, 0.1421, -0.4144, -0.1659, 0.3712 -10.5140]^{\mathrm{T}}$. The norm of parameter vectors are shown in Figs. 1 and 2, where $\|\hat{\theta}_V^{(i)}\|$ and $\|\hat{\theta}_u^{(i)}\|$ converge to 8.1462 and 31.7143 respectively. By using the convergent parameter vector $\hat{\theta}_u^{(20)}$, closed-loop simulation is conducted with the final control policy $\hat{u}^{(20)}$, and Figs. 3 and 5 give the control and state trajectories, respectively. It is indicated from Fig. 3 that the control constraint $|u| \leqslant 0.2$ is satisfied. To show the real cost generated by a control policy u, define

$$J(t) \triangleq \int_0^t Q(x(\tau)) + W(u(\tau))\mathrm{d}\tau.$$

Fig. 4 gives the trajectory of $J(t)$ under the final control policy $\hat{u}^{(20)}$, from which it is observed that $J(t)$ approaches to 0.6781 as time increases.

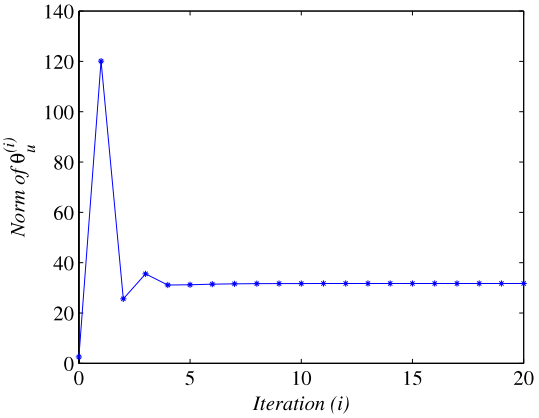

Fig. 2. The norm $\|\hat{\theta}_u^{(i)}\|$ at each iteration.

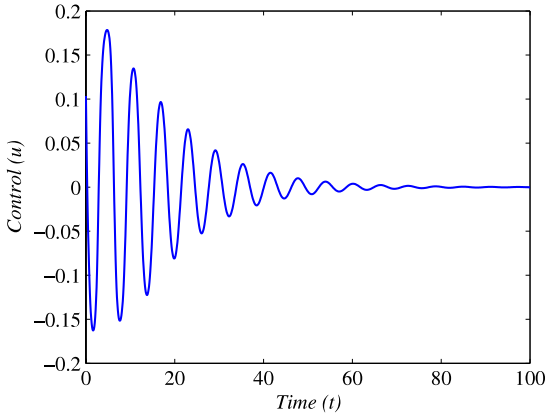

Fig. 3. The final control policy $\hat{u}^{(20)}$.

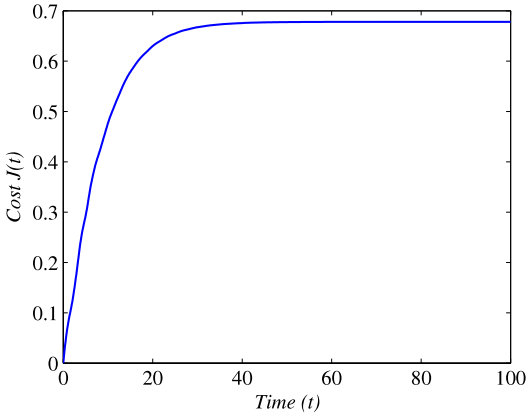

Fig. 4. The cost $J(t)$ with the final control policy $\hat{u}^{(20)}$.

6. Conclusions

An off-policy RL method has been developed for solving the data-based constrained optimal control problem of nonlinear systems, which learns the optimal control policy from real system data rather than mathematical model, and thus avoids the solution of the complicated HJBE. Theoretically, it is found that the off-policy RL method is equivalent to the model-based successive approximation approach for solving the HJBE, and thus its convergence has been proved. To solve the iterative equation in the off-policy RL method, the MWR and the numerically efficient Monte-Carlo integration approaches have been introduced for its implementation. The off-policy RL algorithm is an offline control design procedure, which learns the constrained optimal control policy offline and then is used for real online control purpose after the convergence of the algorithm. Finally, the effectiveness of the developed off-policy RL method has been demonstrated through simulation studies on a rotational/translational actuator system.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61233001, 61273140, 61304086, and 61374105, in part by Beijing Natural Science Foundation under Grant 4132078, in part by the Early Career Development Award of SKLMCCS and in part by the NPRP grant #NPRP 4-1162-1-181 from the Qatar National Research Fund (a member of Qatar Foundation). The authors would like to thank anonymous reviewers for their valuable comments.

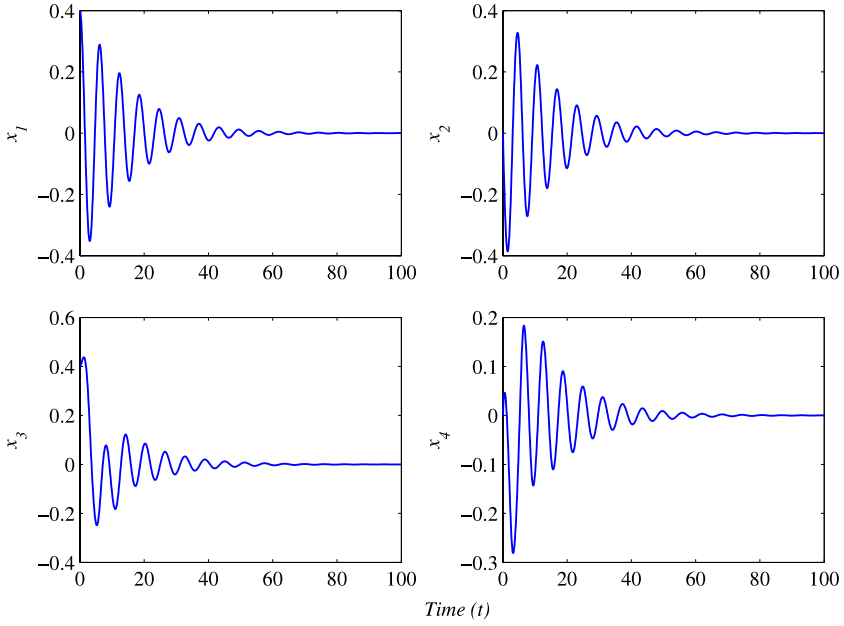

Fig. 5. System state trajectories with the final control policy $\hat{u}^{(20)}$.

References

Abu-Khalaf, M., & Lewis, F. L. (2005). Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, *41*(5), 779–791.

Abu-Khalaf, M., Lewis, F. L., & Huang, J. (2008). Neurodynamic programming and zero-sum games for constrained control systems. *IEEE Transactions on Neural Networks*, *19*(7), 1243–1252.

Anderson, B. D., & Moore, J. B. (2007). *Optimal control: linear quadratic methods*. Mineola, NY: Dover Publications.

Beard, R. W., Saridis, G. N., & Wen, J. T. (1997). Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation. *Automatica*, *33*(12), 2159–2177.

Bertsekas, D. P. (2005). *Dynamic programming and optimal control. Vol. 1*. Nashua: Athena Scientific.

Chen, Z., & Jagannathan, S. (2008). Generalized Hamilton–Jacobi–Bellman formulation-based neural network control of affine nonlinear discrete-time systems. *IEEE Transactions on Neural Networks*, *19*(1), 90–106.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*(1), 219–245.

Faust, A., Ruymgaart, P., Salman, M., Fierro, R., & Tapia, L. (2014). Continuous action reinforcement learning for control-affine systems with unknown dynamics. *IEEE/CAA Journal of Automatica Sinica*, *1*(3), 323–336.

He, P., & Jagannathan, S. (2005). Reinforcement learning-based output feedback control of nonlinear systems with input constraints. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *35*(1), 150–154.

He, P., & Jagannathan, S. (2007). Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *37*(2), 425–436.

Heydari, A., & Balakrishnan, S. N. (2013). Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(1), 147–157.

Hull, D. G. (2003). *Optimal control theory for applications*. Troy, NY: Springer.

Jiang, Y., & Jiang, Z.-P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, *48*(10), 2699–2704.

Jiang, Y., & Jiang, Z.-P. (2014). Robust adaptive dynamic programming and feedback stabilization of nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 882–893.

Lee, J. Y., Park, J. B., & Choi, Y. H. (2012). Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems. *Automatica*, *48*(11), 2850–2859.

Lendaris, G. G. (2009). Adaptive dynamic programming approach to experience-based systems identification and control. *Neural Networks*, *22*(5–6), 822–832.

Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2013). *Optimal control*. Hoboken. NJ: John Wiley & Sons.

Liu, D., Wang, D., & Li, H. (2014). Decentralized stabilization for a class of continuous-time nonlinear interconnected systems using online learning optimal control approach. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(2), 418–428.

Liu, D., Wang, D., & Yang, X. (2013). An iterative adaptive dynamic programming algorithm for optimal control of unknown discrete-time nonlinear systems with constrained inputs. *Information Sciences*, *220*, 331–342.

Liu, D., & Wei, Q. (2014). Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(3), 621–634.

Luo, B., Wu, H.-N., & Huang, T. (2015). Off-policy reinforcement learning for H_∞ control design. *IEEE Transactions on Cybernetics*, *45*(1), 65–76.

Luo, B., Wu, H.-N., Huang, T., & Liu, D. (2014). Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design. *Automatica*, *50*(12), 3281–3290.

Lyashevskiy, S. (1996). Constrained optimization and control of nonlinear systems: new results in optimal control. In *Proceedings of the 35th IEEE decision and control* (pp. 541–546).

Lyshevski, S. E. (1998). Optimal control of nonlinear continuous-time systems: design of bounded controllers via generalized nonquadratic functionals. In *Proceedings of the 1998 American control conference. Vol. 1* (pp. 205–209). IEEE.

Maei, H. R., Szepesvári, C., Bhatnagar, S., & Sutton, R. S. (2010). Toward off-policy learning control with function approximation. In *Proceedings of the 27th international conference on machine learning* (pp. 719–726).

Modares, H., & Lewis, F. L. (2014). Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. *IEEE Transactions on Automatic Control*, *59*(11), 3051–3056.

Modares, H., Lewis, F. L., & Naghibi-Sistani, M.-B. (2013). Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(10), 1513–1525.

Modares, H., Lewis, F. L., & Naghibi-Sistani, M.-B. (2014). Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. *Automatica*, *50*(1), 193–202.

Murray, J. J., Cox, C. J., Lendaris, G. G., & Saeks, R. (2002). Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *32*(2), 140–153.

Peter Lepage, G. (1978). A new algorithm for adaptive multidimensional integration. *Journal of Computational Physics*, *27*(2), 192–203.

Powell, W. B. (2007). *Approximate dynamic programming: solving the curses of dimensionality*. Hoboken, NJ: John Wiley & Sons.

Precup, D., Sutton, R. S., & Dasgupta, S. (2001). Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th international conference on machine learning* (pp. 417–424).

Saridis, G. N., & Lee, C.-S. G. (1979). An approximation theory of optimal control for trainable manipulators. *IEEE Transactions on Systems, Man and Cybernetics*, *9*(3), 152–159.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge: The MIT Press.

Vamvoudakis, K. G., & Lewis, F. L. (2010). Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, *46*(5), 878–888.

Vrabie, D., & Lewis, F. L. (2009). Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, *22*(3), 237–246.

Vrabie, D., Pastravanu, O., Abu-Khalaf, M., & Lewis, F. L. (2009). Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, *45*(2), 477–484.

Wang, D., Liu, D., & Li, H. (2014). Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems. *IEEE Transactions on Automation Science and Engineering*, *11*(2), 627–632.

Wei, Q., & Liu, D. (2012). An iterative ϵ-optimal control scheme for a class of discrete-time nonlinear systems with unfixed initial state. *Neural Networks*, *32*, 236–244.

Yang, X., Liu, D., & Wang, D. (2014). Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints. *International Journal of Control*, *87*(3), 553–566.

Yang, X., Liu, D., Wang, D., & Wei, Q. (2014). Discrete-time online learning control for a class of unknown nonaffine nonlinear systems using reinforcement learning. *Neural Networks*, *55*, 30–41.

Zhang, H., Luo, Y., & Liu, D. (2009). Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. *IEEE Transactions on Neural Networks*, *20*(9), 1490–1503.

Zhao, Q., Xu, H., & Jagannathan, S. (2014). Near optimal output feedback control of nonlinear discrete-time systems based on reinforcement neural network learning. *IEEE/CAA Journal of Automatica Sinica*, *1*(4), 372–384.