Neural Comput & Applic (2013) 23:1843–1850
DOI 10.1007/s00521-012-1249-y

ISNN2012

Adaptive optimal control for a class of continuous-time affine nonlinear systems with unknown internal dynamics

Derong Liu · Xiong Yang · Hongliang Li

Received: 3 April 2012 / Accepted: 20 October 2012 / Published online: 15 November 2012

Abstract This paper develops an online algorithm based on policy iteration for optimal control with infinite horizon cost for continuous-time nonlinear systems. In the present method, a discounted value function is employed, which is considered to be a more general case for optimal control problems. Meanwhile, without knowledge of the internal system dynamics, the algorithm can converge uniformly online to the optimal control, which is the solution of the modified Hamilton–Jacobi–Bellman equation. By means of two neural networks, the algorithm is able to find suitable approximations of both the optimal control and the optimal cost. The uniform convergence to the optimal control is shown, guaranteeing the stability of the nonlinear system. A simulation example is provided to illustrate the effectiveness and applicability of the present approach.

Keywords Adaptive dynamic programming · Reinforcement learning · Policy iteration · Adaptive optimal control · Neural network · Online control · Nonlinear system

1 Introduction

Optimal control has drawn considerable attention since it was formally developed about five decades ago by Pontryagin [1] and Bellman [2]. Nowadays, with the development of the theory, optimal control has been one of the fundamental principles of modern control systems design. From mathematical perspectives, the solutions of control problems can be obtained by solving the Hamilton–Jacobi–Bellman (HJB) equation, which guarantees the sufficient conditions for existence of optimality [3]. For linear dynamic systems and quadratic costs, the HJB equation reduces to the Riccati equation, which can be accurately solved by analytical or numerical methods [4–6]. However, in the case of continuous-time (CT) nonlinear systems, the HJB equation is actually a nonlinear partial differential equation, which is extremely intractable to solve by analytical approaches. Consequently, a huge number of significant efforts [7–10] have been made to develop algorithms which approximately solve this type of equations. Furthermore, large amounts of important measures to deal with discrete-time (DT) HJB equations [11–15] have also been taken.

Among the methods involved in computational intelligence techniques, there are two approaches known as value iteration (VI) [11] and policy iteration (PI) [16]. The main difference between PI and VI is that PI requires a stabilizing initial policy while VI does not, but VI cannot guarantee the stability of control policies derived during iteration at each step. In [17, 18], both of the authors proposed an online PI algorithm for CT nonlinear systems which converges to the optimal control. However, in [17], only partial knowledge of the nonlinear system was required. After that, Lee et al. [19] developed an online algorithm based on generalized VI technique for uncertain CT linear systems with the performance index involving a discount factor.

Motivated by the above work, in this paper, we investigate an online approximate optimal control based on PI

D. Liu (✉) · X. Yang · H. Li
State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
e-mail: derongliu@gmail.com

X. Yang
e-mail: yangxiong1983@126.com

H. Li
e-mail: hongliang.li@ia.ac.cn

for partially unknown CT nonlinear systems with the infinite horizon cost involving a discount factor. As we know, for DT nonlinear systems, there is no distinct difference for HJB equations [14, 15], whether the discount factor is involved in the performance index or not. However, there are distinctly different characteristics between the present HJB equation (we call it the modified HJB equation in this paper) and the classical HJB equation for CT nonlinear systems. To the best of our knowledge, the discount factor, which was generally setting in optimal control problems, was seldom considered for CT nonlinear dynamic systems. Moreover, there is no strict proof of the convergence to the optimal control which is the solution of the modified HJB equation. Consequently, a significant difference between [9, 17–19] and the present investigation is that, in our case, the convergence to the optimal control with the infinite horizon cost involving a discount factor is explicitly established. Meanwhile, we show that the performance of the convergence to the optimal control is closely linked with the discount factor in the simulation example. Furthermore, this paper can be viewed as the generalized form of the literatures [9, 13, 17, 18].

The rest of this paper is arranged as follows. Section 2 presents preliminaries of optimal control problems for nonlinear systems. Section 3 gives an online algorithm based on PI to solve the modified HJB equation, and shows the convergence of the algorithm. Section 4 provides the formulation of the algorithm by using neural networks approximation. Section 5 presents a simulation example to complete the theoretical discussions. Finally, in Sect. 6, the paper is concluded with several remarks.

2 Preliminaries of optimal control problems

For purpose of the present paper, we consider a time-invariant input affine plant of the form

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

where $x(t) \in \mathbb{R}^n, f(x) \in \mathbb{R}^n, g(x) \in \mathbb{R}^{n\times m}$ and $u(x) \in \mathbb{R}^m$. Assume that $f(0) = 0$ and $f(x) + g(x)u$ is Lipschitz continuous on a compact set $\Omega \subseteq \mathbb{R}^n$ which contains the origin. The nonlinear system is stabilizable on Ω, that is, there exists a control policy $u(t)$ such that the given system is asymptotically stable on Ω.

Definition 1 (Discount factor for CT systems [20]) Let, $\lambda > 0$ and $r(x,u) \geq 0$. If the performance index for a CT system $\dot{x} = f(x, u)$ has the form

$$V(x(0)) = \int_0^\infty e^{\lambda\tau} r(x(\tau), u(\tau))\mathrm{d}\tau,$$

then λ is the discount factor, and $V(x)$ is the discounted value function for the CT system.

In this paper, a discounted value function for system (1) is described by

$$V(x(t)) = \int_t^\infty e^{\alpha(\tau-t)} r(x(\tau), u(\tau))\mathrm{d}\tau,\ (\tau \geq t), \tag{2}$$

where $\alpha > 0$, $r(x,u) = Q(x) + u^{\mathsf{T}}Ru$ (T is the transposition symbol), and $Q(x)$ is positive definite, that is, for $\forall\ x \neq 0$, $Q(x) > 0$ and $x = 0 \Leftrightarrow Q(x) = 0$. Furthermore, R is a symmetric positive definite matrix.

Definition 2 (Admissible control [8]) A control $u(x)$: $\mathbb{R}^n \to \mathbb{R}^m$ is defined to be admissible with respect to (2) on Ω, written as $u(x) \in \mathscr{A}(\Omega)$, if $u(x)$ is continuous on $\Omega, u(0) = 0, u(x)$ stabilizes system (1) on Ω and $V(x_0)$ is finite for every $x \in \Omega$.

Given a control $u(x) \in \mathscr{A}(\Omega)$, if the associated value function $V(x) \in \mathrm{C}^1(\Omega)$, then its infinitesimal version is

$$\begin{aligned} &(V_x)^{\mathsf{T}}(f(x) + g(x)u(x)) + \alpha V(x) + r(x, u(x)) = 0, \\ &V(0) = 0, \end{aligned} \tag{3}$$

where V_x denotes the partial derivative of the value function $V(x(t))$ with respect to x. Actually, (3) is the generalized Hamilton–Jacobi–Bellman (GHJB) equation for the nonlinear system [8].

Define the pre-Hamilton function for the control $u(x) \in \mathscr{A}(\Omega)$ and the associated value function $V(x)$ by

$$\begin{aligned} H(x, V_x, u) = {} & (V_x)^{\mathsf{T}}(f(x) + g(x)u(x)) \\ & + \alpha V(x) + r(x, u(x)). \end{aligned}$$

Then, the optimal cost $V^*(x)$ is obtained by solving the HJB equation that

$$\min_{u(x)\in\mathscr{A}(\Omega)} H(x, V_x^*, u) = 0. \tag{4}$$

Suppose that the minimum value on the left-hand side of the Eq. (4) exists and is unique. Then, the optimal control for system (1) with the performance index (2) is

$$u^*(x) = -\frac{1}{2}R^{-1}g^{\mathsf{T}}(x)V_x^*. \tag{5}$$

Substituting (5) into (3), we derive the modified HJB equation as

$$\begin{aligned} &(V_x^*)^{\mathsf{T}} f(x) + \alpha V^*(x) + Q(x) \\ &\quad - \frac{1}{4}(V_x^*)^{\mathsf{T}} g(x) R^{-1} g^{\mathsf{T}}(x) V_x^* = 0, \\ &V^*(0) = 0. \end{aligned} \tag{6}$$

Remark 1 From (6), one shall note that there is α $V^*(x)$ involved in the HJB equation. To solve (6), we need

the knowledge of both αV^* and V_x^*. However, this is quite different from the form appeared in classic literatures [7–9]. Consequently, we call (6) the modified HJB equation.

3 Policy iteration algorithm for solving optimal control problems

The purpose of this section is devoted to present an online iterative algorithm for solving optimal control problems. And the convergence of the algorithm to the optimal control is developed.

(a) Given an initial control $u^{(i)}(x) \in \mathscr{A}(\Omega)$, solve the value function $V^{(i)}(x)$ by

$$\begin{aligned} V^{(i)}(x(t)) &= \int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau)) \mathrm{d}\tau \\ &\quad + e^{\alpha T} V^{(i)}(x(t+T)), \\ V^{(i)}(0) &= 0. \end{aligned} \tag{7}$$

(b) Update the control policy via

$$\begin{aligned} u^{(i+1)}(x) &= \min_{u \in \mathscr{A}(\Omega)} H(x, V_x^{(i)}, u) \\ &= -\frac{1}{2} R^{-1} g^{\mathsf{T}}(x) V_x^{(i)}. \end{aligned} \tag{8}$$

From (7) and (8), one shall notice that the iteration algorithm can solve the optimal control problem without knowledge of the internal dynamics of the nonlinear system, that is, the knowledge of $f(x)$ is not required. This online policy iteration algorithm was first proposed in [17]. In this paper, we apply this method to investigate CT nonlinear systems with a discount factor in the value function.

In fact, (7) can be viewed as the GHJB equation for the nonlinear system. For convenience, we denote it as follows:

$$\begin{aligned} \mathrm{GHJB}(V^{(i)}, u^{(i)}) &= \int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau)) \mathrm{d}\tau \\ &\quad + e^{\alpha T} V^{(i)}(x(t+T)) - V^{(i)}(x(t)) \\ &= 0, \\ V^{(i)}(0) &= 0. \end{aligned} \tag{9}$$

Prior to showing the convergence of the iteration algorithm, we need to establish the following lemma for our further discussion.

Lemma 1 *Assume that $u^{(i)}(x) \in \mathscr{A}(\Omega)$; then, $V^{(i)}(x)$ is the unique solution of* (9) *if and only if $V^{(i)}(x)$ is the only solution of*

$$\begin{aligned} &(V_x^{(i)})^{\mathsf{T}}(f(x) + g(x)u^{(i)}(x)) + \alpha V^{(i)}(x) + r(x, u^{(i)}(x)) = 0, \\ &V^{(i)}(0) = 0. \end{aligned} \tag{10}$$

Proof We divide the proof into two parts: (i) we show the solution of (9) is equivalent to the solution of (10); (ii) we show that there exists a unique solution of (9).

(i) (Sufficiency) Since $V^{(i)}(x)$ is the solution of (9), we have that

$$\begin{aligned} &\lim_{T\to 0} \frac{1}{T} \int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau)) \mathrm{d}\tau \\ &= \lim_{T\to 0} \frac{1}{T} \left[V^{(i)}(x(t)) - e^{\alpha T} V^{(i)}(x(t+T)) \right] \\ &= \lim_{T\to 0} \frac{1}{T} \left[V^{(i)}(x(t)) - V^{(i)}(x(t+T)) \right] \\ &\quad + \lim_{T\to 0} \frac{1 - e^{\alpha T}}{T} V^{(i)}(x(t+T)). \end{aligned}$$

That is,

$$\begin{aligned} r(x, u^{(i)}(x)) &= -\dot{V}^{(i)}(x(t)) - \alpha V^{(i)}(x(t)) \\ &= -(V_x^{(i)})^{\mathsf{T}}(f(x) + g(x)u^{(i)}) - \alpha V^{(i)}(x). \end{aligned}$$

Accordingly, $V^{(i)}(x)$ is the solution of (10). (Necessity) Suppose that $V^{(i)}(x)$ is the solution of (10). Then, we get

$$\begin{aligned} -e^{\alpha t} r(x, u^{(i)}(x)) &= e^{\alpha t} \left[(V_x^{(i)})^{\mathsf{T}}(f(x) + g(x)u^{(i)}) \right. \\ &\quad \left. + \alpha V^{(i)}(x) \right] \\ &= \frac{\mathrm{d}(e^{\alpha t} V^{(i)} x(t))}{\mathrm{d}t}. \end{aligned} \tag{11}$$

Integrating both sides of (11) over the time interval $[t, t+T]$, we obtain that

$$\begin{aligned} &-\int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau)) \mathrm{d}\tau \\ &= e^{\alpha T} V^{(i)}(x(t+T)) - V^{(i)}(x(t)). \end{aligned}$$

Consequently, $V^{(i)}(x)$ is the solution of (9).

(ii) In light of $u^{(i)}(x) \in \mathscr{A}(\Omega)$, we have that $V^{(i)}(x) \in \mathrm{C}^1(\Omega)$. Let, $h(V^{(i)}, t) = \alpha V^{(i)}(x(t)) + r(x(t), u^{(i)}(t))$. Then, for $\forall V_1^{(i)} \in V(\Omega)$, $V_2^{(i)} \in V(\Omega)$, we obtain that

$$|h(V_1^{(i)}, t) - h(V_2^{(i)}, t)| = \alpha |V_1^{(i)}(x(t)) - V_2^{(i)}(x(t))|.$$

Obviously, $h(V^{(i)}, t)$ satisfies the Lipschitz condition on $V^{(i)}(\Omega) \times [0, \infty)$. By means of the theory of ordinary differential equations (Picard theorem), we can derive that (10) has the unique solution. Considering part (i), there exists the unique solution of (9). □

Theorem 1 *Let $u^{(i)}(x) \in \mathscr{A}(\Omega)$ and $V^{(i)}(x) \in \mathrm{C}^1(\Omega)$ satisfies* $\mathrm{GHJB}(V^{(i)}, u^{(i)}) = 0$ *with the boundary condition*

$V^{(i)}(0) = 0$. *Then, the control policy* (8) *is an admissible control for the system* (1) *on* Ω. *Furthermore, if* $V^{(i+1)}(x)$ *is the unique positive definite function such that* $\mathrm{GHJB}(V^{(i+1)}, u^{(i+1)}) = 0$ *with* $V^{(i+1)}(0) = 0$, *then* $V^*(x) \le V^{(i+1)}(x) \le V^{(i)}(x)$.

Proof We divide the proof into two parts. First, we show that $u^{(i+1)}(x) \in \mathscr{A}(\Omega)$. Second, we show that the sequence $\{V^{(i)}(x)\}$ is monotonically nonincreasing over the compact set Ω, that is, $V^*(x) \le V^{(i+1)}(x) \le V^{(i)}(x)$.

(*a*) The proof for $u^{(i+1)}(x) \in \mathscr{A}(\Omega)$.

Observing the expression of $u^{(i+1)}(x)$ in (8) and $V^{(i)} \in \mathrm{C}^1(\Omega)$, we obtain that $u^{(i+1)}(x) \in \mathrm{C}(\Omega)$.

In light of $V^{(i)}(x(t)) \ge 0$ and $V^{(i)}(x(t)) = 0 \Leftrightarrow x(t) = 0$, we have that $V_x^{(i)}|_{x=0} = 0$. Hence, $u^{(i+1)}(0) = 0$. Taking the derivative of $V^{(i)}(x)$ along the trajectory of system $f(x) + g(x)u^{(i+1)}$, we get that

$$\dot{V}^{(i)}(x, u^{(i+1)}) = (V_x^{(i)})^{\mathsf{T}} f + (V_x^{(i)})^{\mathsf{T}} g u^{(i+1)}. \tag{12}$$

Since $\mathrm{GHJB}(V^{(i)}, u^{(i)}) = 0$, by using Lemma 1, we have that

$$(V_x^{(i)})^{\mathsf{T}} f = -(V_x^{(i)})^{\mathsf{T}} g u^{(i)} - \alpha V^{(i)} - r(x, u^{(i)}).$$

Therefore, (12) can be rewritten as

$$\begin{aligned}\dot{V}^{(i)}(x, u^{(i+1)}) = &-(V_x^{(i)})^{\mathsf{T}} g u^{(i)} + (V_x^{(i)})^{\mathsf{T}} g u^{(i+1)} \\ &- \alpha V^{(i)} - r(x, u^{(i)}).\end{aligned} \tag{13}$$

From (8) and (13), we obtain that

$$\begin{aligned}\dot{V}^{(i)}(x, u^{(i+1)}) = &-\alpha V^{(i)} - Q(x) - \Big[(u^{(i)})^{\mathsf{T}} R u^{(i)} \\ &+ 2(u^{(i+1)})^{\mathsf{T}} R (u^{(i+1)} - u^{(i)})\Big].\end{aligned} \tag{14}$$

Since R is a positive definite matrix, we can denote $R = \Lambda\Sigma\Lambda$, where Λ is an orthogonal symmetric matrix, and Σ is the diagonal matrix with its values being the singular values of R. Denote that $\Sigma = (\Sigma_{kk})$. Then, we have that $\Sigma_{kk} > 0$, $k = 1, 2, \ldots, n$. Moreover, change the coordinate with $u^{(i)} = \Lambda^{-1} z^{(i)}$. Let,

$$\pounds = 2(u^{(i+1)})^{\mathsf{T}} R (u^{(i+1)} - u^{(i)}) + (u^{(i)})^{\mathsf{T}} R u^{(i)}.$$

Then, we derive that

$$\begin{aligned}\pounds &= 2(z^{(i+1)})^{\mathsf{T}} \Lambda^{-1} (\Lambda\Sigma\Lambda)(\Lambda^{-1} z^{(i+1)} - \Lambda^{-1} z^{(i)}) \\ &\quad + (z^{(i)})^{\mathsf{T}} \Lambda^{-1} (\Lambda\Sigma\Lambda) \Lambda^{-1} z^{(i)} \\ &= 2(z^{(i+1)})^{\mathsf{T}} \Sigma (z^{(i+1)} - z^{(i)}) + (z^{(i)})^{\mathsf{T}} \Sigma (z^{(i)}) \\ &= \sum_{k=1}^{m} \Sigma_{kk} \Big[2 z_k^{(i+1)} (z_k^{(i+1)} - z_k^{(i)}) + (z_k^{(i)})^2\Big] \\ &= \sum_{k=1}^{m} \Sigma_{kk} \Big[(z_k^{(i+1)})^2 + (z_k^{(i+1)} - z_k^{(i)})^2\Big] \ge 0.\end{aligned} \tag{15}$$

Noticing that $\alpha V^{(i)}(x) > 0$ and $Q(x) > 0$ for $\forall x \ne 0$, by means of (14) and (15), we obtain that $\dot{V}^{(i)}(x, u^{(i+1)}) < 0$. Therefore, $V^{(i)}(x)$ is the Lyapunov function for $u^{(i+1)}(x)$ on Ω. Hence, we derive that $u^{(i+1)}(x) \in \mathscr{A}(\Omega)$.

(*b*) The proof for $V^*(x) \le V^{(i+1)}(x) \le V^{(i)}(x)$.

Taking the derivative of $V^{(i)}(x)$ along the trajectory of system $f + gu^{(i+1)}$, we have that

$$\begin{aligned}&V^{(i+1)}(x(t)) - V^{(i)}(x(t)) \\ &= -\int_t^{\infty} \frac{\mathrm{d}(V^{(i+1)} - V^{(i)})}{\mathrm{d}x} (f + g u^{(i+1)}) \mathrm{d}\tau.\end{aligned} \tag{16}$$

Observing that $\mathrm{GHJB}(V^{(i)}, u^{i}) = 0$, $\mathrm{GHJB}(V^{(i+1)}, u^{i+1}) = 0$, we obtain that

$$(V_x^{(i)})^{\mathsf{T}} f = -(V_x^{(i)})^{\mathsf{T}} g u^{(i)} - \alpha V^{(i)} - r(x, u^{(i)}), \tag{17}$$

$$\begin{aligned}(V_x^{(i+1)})^{\mathsf{T}} f = &-(V_x^{(i+1)})^{\mathsf{T}} g u^{(i+1)} \\ &- \alpha V^{(i+1)} - r(x, u^{(i+1)}).\end{aligned} \tag{18}$$

By virtue of (16)–(18), we derive that

$$\begin{aligned}&V^{(i+1)}(x(t)) - V^{(i)}(x(t)) \\ &\quad = \int_t^{\infty} \Big\{\alpha[V^{(i+1)}(x(\tau)) - V^{(i)}(x(\tau))] - \mathscr{R}(u(\tau))\Big\} \mathrm{d}\tau,\end{aligned} \tag{19}$$

where

$$\begin{aligned}\mathscr{R}(u) = &\, 2(u^{(i+1)})^{\mathsf{T}} R (u^{(i+1)} - u^{(i)}) \\ &+ (u^{(i)})^{\mathsf{T}} R u^{(i)} - (u^{(i+1)})^{\mathsf{T}} R u^{(i+1)}.\end{aligned}$$

By the same technique used in part (*a*), we can prove that

$$\mathscr{R}(u(t)) = \sum_{k=1}^{m} \Sigma_{kk} (z_k^{(i+1)} - z_k^{(i)})^2 \ge 0.$$

Let, $F(t) = V^{(i+1)}(x(t)) - V^{(i)}(x(t))$; then, (19) can be rewritten as

$$F(t) = \alpha \int_t^{\infty} F(\tau) \mathrm{d}\tau - \int_t^{\infty} \mathscr{R}(u(\tau)) \mathrm{d}\tau. \tag{20}$$

Taking the derivative of both sides of (20) with respect to t, we get that

$$\dot{F}(t) + \alpha F(t) = \mathscr{R}(u(t)) \ge 0. \tag{21}$$

Multiplying $e^{\alpha t}$ to both sides of (21), we can derive that

$$\frac{\mathrm{d}(e^{\alpha t} F(t))}{\mathrm{d}t} \ge 0. \tag{22}$$

Integrating both sides of (22) over the time interval $[t, \infty)$, and noticing that $F(\infty) = 0$, we obtain that $F(t) \le 0$. Therefore, $V^{(i+1)}(x(t)) \le V^{(i)}(x(t))$. Furthermore, it can be shown by contradiction that

$V^*(x(t)) \le V^{(i+1)}(x(t)), \forall\, t \ge 0.$

Let, $f(x) = (f^1(x), f^2(x), \ldots, f^n(x)) \in \mathbb{R}^n$, and define the norm of the vector-valued function $f(x)$ as $\|f\| = \sup_{x\in\Omega,\, 1\le k\le n}\{|f^k(x)|\}$.

Definition 3 (Uniform Convergence) A sequence of vector-valued function of $f_m(x) \in \mathbb{R}^n$ is said to converge uniformly to $f(x)$ with the norm $\|\cdot\|$ on a set Ω if, for $\forall\, \varepsilon > 0$, there exists a positive N (depending only on ε) such that $m > N$ implies $\|f_m(x) - f(x)\| < \varepsilon$. For brief, we write that, for $m \to \infty$, $f_m(x) \rightrightarrows f(x)$, $\forall\, x \in \Omega$.

Theorem 2 *Given an initial control $u^{(0)}(x) \in \mathscr{A}(\Omega)$, then $u^{(i)}(x) \in \mathscr{A}(\Omega)$, $\forall\, i \ge 0$. Furthermore, for $\forall\, \varepsilon > 0$, there exists $i_0 \in \mathbb{N}$ such that $i \ge i_0$ implies*

$$\sup_{x\in\Omega} |V^{(i)}(x) - V^*(x)| < \varepsilon, \|u^{(i)}(x) - u^*(x)\| < \varepsilon.$$

Proof By means of Theorem 1 and the method of mathematical induction, we have that $u^{(i)}(x) \in \mathscr{A}(\Omega)$, $\forall\, i \ge 0$. Meanwhile, we obtain the decreasing sequence

$$V^{(1)}(x) \ge V^{(2)}(x) \ge \cdots \ge V^{(i+1)}(x) \ge \cdots \ge V^*(x).$$

Therefore, by employing the monotone convergence theorem [21], we get that, for each fixed $\tilde{x} \in \Omega$, there exists $\lim_{i\to\infty} V^{(i)}(\tilde{x}) = \inf_{i\in\mathbb{N}}\{V^{(i)}(\tilde{x})\}$. In view of $\inf_{i\in\mathbb{N}}\{V^{(i)}(\tilde{x})\} = V^*(\tilde{x})$, we obtain $\lim_{i\to\infty} V^{(i)}(\tilde{x}) = V^*(\tilde{x})$. Since $\Omega \subset \mathbb{R}^n$ is a compact set, by virtue of Dini's theorem, we can derive that, for $i \to \infty$,

$$V^{(i)}(x) \rightrightarrows V^*(x), \forall\, x \in \Omega.$$

Observing that $V^{(i)}(x) \in \mathbb{R}$ and by the definition of the norm $\|\cdot\|$, we obtain that, for $\forall \varepsilon > 0$, there exists $i_1 \in \mathbb{N}$ such that $i \ge i_1$ implies

$$\sup_{x\in\Omega} |V^{(i)}(x) - V^*(x)| < \varepsilon.$$

Meanwhile, it implies that the sequence of system trajectories is uniformly convergent. Consequently, $u^{(i)}$ is also uniformly convergent on Ω, that is, $V_x^{(i)}$ is uniformly convergent on Ω. Noticing that $V^{(i)}(x) \in \mathrm{C}^1(\Omega)$, by employing the theorem about the relationship between the uniform convergence and differentiation [21], we derive that, for $i \to \infty$, $V_x^{(i)} \rightrightarrows V_x^*$. Therefore, for $i \to \infty$, there exists

$$u^{(i)}(x) \rightrightarrows u^*(x), \forall\, x \in \Omega.$$

In view of $u^{(i)}(x) \in \mathbb{R}^n$, we have that, for $\forall\, \varepsilon > 0$, there exists $i_2 \in \mathbb{N}$ such that $i \ge i_2$ implies $\|u^{(i)}(x) - u^*(x)\| < \varepsilon$. Let, $i_0 = \max\{i_1, i_2\}$. Accordingly, for $\forall\, \varepsilon > 0$, there exists $i_0 \in \mathbb{N}$ such that $i \ge i_0$ implies

$$\sup_{x\in\Omega} |V^{(i)}(x) - V^*(x)| < \varepsilon, \|u^{(i)}(x) - u^*(x)\| < \varepsilon.$$

From Theorem 2, one can draw the conclusion that the proposed online policy iteration algorithm in (7) and (8) shall converge to the solution of optimal control problems (1) and (2). Therefore, in order to derive the optimal control, one does not need to have the knowledge of $f(x)$.

4 Neural network-based least-squares approximate HJB solution

In this section, neural networks (NNs) are used to solve approximately for the value function $V^{(i)}(x)$ with arbitrary $x \in \Omega$ in (7). It is well known that NNs are able to approximate smooth time-invariant functions on given compact sets [22]. Hence, one can approximate the value function $V^{(i)}(x)$ on the compact set Ω by

$$V_L^{(i)}(x) = \sum_{i=1}^{L} \omega_j^{(i)} \sigma_j(x) = (\boldsymbol{\omega}_L^{(i)})^{\mathsf{T}} \boldsymbol{\sigma}_L(x), \tag{23}$$

where $\boldsymbol{\omega}_L^{(i)} = [\omega_1^{(i)}, \omega_2^{(i)}, \ldots, \omega_L^{(i)}]^{\mathsf{T}}$ is the weight vector, L is the number of neurons in the hidden layers, $\boldsymbol{\sigma}_L(x) = [\sigma_1(x), \sigma_2(x), \ldots, \sigma_L(x)]^{\mathsf{T}}$ is the vector activation function, and $\sigma_j(x) \in \mathrm{C}^1(\Omega)$, $\sigma_j(0) = 0$. The set $\{\sigma_j(x)\}_1^L$ is often selected to be linearly independent.

By replacing $V^{(i)}(x)$ with $V_L^{(i)}(x)$ in (7), we have that

$$\begin{aligned}(\boldsymbol{\omega}_L^{(i)})^{\mathsf{T}} \boldsymbol{\sigma}_L(x(t)) = &\int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau))\mathrm{d}\tau \\ &+ e^{\alpha T} (\boldsymbol{\omega}_L^{(i)})^{\mathsf{T}} \boldsymbol{\sigma}_L(x(t+T)),\end{aligned} \tag{24}$$

and the residual error is

$$\begin{aligned}e_L^{(i)}(x, T) = &\int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau))\mathrm{d}\tau \\ &+ (\boldsymbol{\omega}_L^{(i)})^{\mathsf{T}} \left[e^{\alpha T} \boldsymbol{\sigma}_L(x(t+T)) - \boldsymbol{\sigma}_L(x(t))\right].\end{aligned} \tag{25}$$

In order to derive the least-squares solution, the method of weighted residual [23] is employed. The weight $\boldsymbol{\omega}_L^{(i)}(x)$ is determined by projecting the residual error onto the term $\mathrm{d}e_L^{(i)}(x, T)/\mathrm{d}\boldsymbol{\omega}_L^{(i)}$ to obtain

$$\left\langle \frac{\mathrm{d}e_L^{(i)}(x, T)}{\mathrm{d}\boldsymbol{\omega}_L^{(i)}}, e_L^{(i)}(x, T) \right\rangle_{\Omega} = 0, \tag{26}$$

where $\langle f, g\rangle_\Omega = \int_\Omega f g^{\mathsf{T}} \mathrm{d}x$ is the Lebesgue integral on Ω. Let

$$\mathfrak{N}(x, T) = e^{\alpha T} \boldsymbol{\sigma}_L(x(t+T)) - \boldsymbol{\sigma}_L(x(t)). \tag{27}$$

From (25)–(27), we derive that

$$\left\langle \mathfrak{N}(x,T), \int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau))\mathrm{d}\tau \right\rangle_{\Omega} + \langle \mathfrak{N}(x,T), \mathfrak{N}(x,T)\rangle_{\Omega} \boldsymbol{\omega}_L^{(i)} = 0. \tag{28}$$

Lemma 2 *If the set* $\{\sigma_j(x)\}_1^L$ *is linearly independent, then the following set*

$$\left\{\nabla\sigma_j^{\mathsf{T}}(f+gu) + \alpha\sigma_j\right\}_1^L$$

is also linearly independent.

Proof If there exists a nonzero $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_L)^{\mathsf{T}} \in R^L$, such that

$$\begin{aligned} &\sum_{j=1}^{L} \lambda_j \left[\nabla\sigma_j^{\mathsf{T}}(f+gu) + \alpha\sigma_j\right] = 0, \\ &\Leftrightarrow e^{\alpha t}\boldsymbol{\lambda}^{\mathsf{T}}\left[\nabla\boldsymbol{\sigma}_L^{\mathsf{T}}(f+gu) + \alpha\boldsymbol{\sigma}_L\right] = 0, \end{aligned} \tag{29}$$

then along the trajectories $x(t; x_0, u)$, we have that

$$\begin{aligned} &\int_t^{\infty} e^{\alpha\tau}\boldsymbol{\lambda}^{\mathsf{T}}\left[\nabla\boldsymbol{\sigma}_L^{\mathsf{T}}(f+gu) + \alpha\boldsymbol{\sigma}_L\right]\mathrm{d}\tau = 0, \\ &\Leftrightarrow \boldsymbol{\lambda}^{\mathsf{T}} \int_t^{\infty} \frac{\mathrm{d}(e^{\alpha\tau}\boldsymbol{\sigma}_L(x(\tau)))}{\mathrm{d}\tau}\mathrm{d}\tau = 0. \end{aligned} \tag{30}$$

Observing that $\boldsymbol{\sigma}_L(x(\infty)) = 0$, by virtue of (30), we have that $\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{\sigma}_L(x) = 0$. Since the set $\{\sigma_j(x)\}_1^L$ is linearly independent, one can conclude $\boldsymbol{\lambda} = 0$, which contradicts $\boldsymbol{\lambda} \neq 0$. Therefore, the set $\left\{\nabla\sigma_j^{\mathsf{T}}(f+gu) + \alpha\sigma_j\right\}_1^L$ is linearly independent. □

Lemma 3 *Assume that* $u(x) \in \mathscr{A}(\Omega)$. *If the set* $\{\sigma_j(x)\}_1^L$ *is linearly independent, then* $\exists\ T > 0$ *such that, for every* $x \in \Omega \setminus \{0\}$, $\{\mathfrak{N}_j(x,T) = e^{\alpha T}\sigma_j(x(t+T)) - \sigma_j(x(t))\}_1^L$ *is also linearly independent.*

Proof In view of $u(x) \in \mathscr{A}(\Omega)$, we can derive that the vector field $f + gu$ is asymptotically stable. Then, along the trajectories $\theta(t; x_0, u)$, $x_0 \in \Omega$, we have that

$$\begin{aligned} \mathfrak{N}_j(x,T) &= \int_t^{t+T} \frac{\mathrm{d}(e^{\alpha(\tau-t)}\sigma_j)}{\mathrm{d}\tau}\mathrm{d}\tau \\ &= \int_t^{t+T} e^{\alpha(\tau-t)}[\alpha\sigma_j + \nabla\sigma_j^{\mathsf{T}}(f+gu)] \\ &\quad \times (\theta(t; x_0, u))\mathrm{d}\tau. \end{aligned}$$

Now, assume that the lemma is not true. Then, there exists a nonzero $\boldsymbol{\kappa} \in \mathbb{R}^n$, for $\forall T > 0$, such that $\boldsymbol{\kappa}^{\mathsf{T}}\mathfrak{N}(x,T) = 0$. This implies that for $\forall x_0$, $\forall T > 0$,

$$\begin{aligned} &\int_t^{t+T} e^{\alpha(\tau-t)}\boldsymbol{\kappa}^{\mathsf{T}}[\alpha\boldsymbol{\sigma}_L + \nabla\boldsymbol{\sigma}_L{}^{\mathsf{T}}(f+gu)]\mathrm{d}\tau = 0, \\ &\Rightarrow \boldsymbol{\kappa}^{\mathsf{T}}[\alpha\boldsymbol{\sigma}_L + \nabla\boldsymbol{\sigma}_L{}^{\mathsf{T}}(f+gu)] = 0, \end{aligned}$$

which contradicts the linear independence of $\{\nabla\sigma_j^{\mathsf{T}}(f+gu) + \alpha\sigma_j\}_1^L$.

By Lemma 3, we know that $\langle \mathfrak{N}(x,T), \mathfrak{N}(x,T)\rangle_{\Omega}$ is invertible. Accordingly, from (28), we can derive that

$$\boldsymbol{\omega}_L^{(i)} = -\langle \mathfrak{N}(x,T), \mathfrak{N}(x,T)\rangle_{\Omega}^{-1} \times \left\langle \mathfrak{N}(x,T), \int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau))\mathrm{d}\tau \right\rangle_{\Omega}. \tag{31}$$

Since the weight $\boldsymbol{\omega}_L^{(i)}$ is available, one can get the improved control policy by

$$u^{(i+1)}(x) = -\frac{1}{2}R^{-1}g^{\mathsf{T}}\nabla\boldsymbol{\sigma}_L^{\mathsf{T}}(x)\boldsymbol{\omega}_L^{(i)}. \tag{32}$$

And (32) can be viewed as the output of the action NN. Therefore, (7) and (8) can be solved by using the critic NN described by (23) and the action NN developed by (32), respectively.

4.1 Algorithm based on NN approximation

Solving the integration in (31) is rather complicated, since the evolution of the L_2 inner product over Ω is required. However, by means of the method in [9], the integral can be well approximated by discretization. A mesh of points of size of δx over the integration region is introduced on Ω. Then, we can define

$$M = \left[\mathfrak{N}(x,T)|_{x_1} \ \cdots \ \mathfrak{N}(x,T)|_{x_p}\right]^{\mathsf{T}},$$

where $\mathfrak{N}(x,T) = e^{\alpha T}\boldsymbol{\sigma}_L(x(t+T)) - \boldsymbol{\sigma}_L(x(t))$, and

$$N = \left[\int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau))\mathrm{d}\tau|_{x_1} \ \cdots \ \int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau))\mathrm{d}\tau|_{x_p}\right]^{\mathsf{T}},$$

where p represents the number of points in the mesh on Ω. Reducing the size of the mesh, we obtain that

$$\langle \mathfrak{N}(x,T), \mathfrak{N}(x,T)\rangle_{\Omega} = \lim_{\|\delta x\|\to 0}(M^{\mathsf{T}}M)\delta x,$$

$$\left\langle \mathfrak{N}(x,T), \int_t^{t+T} e^{\alpha(\tau-t)} r(x(\tau), u^{(i)}(\tau))\mathrm{d}\tau \right\rangle_{\Omega} = \lim_{\|\delta x\|\to 0}(M^{\mathsf{T}}N)\delta x.$$

Therefore, (31) can be rewritten as

$$\boldsymbol{\omega}_{L,p}^{(i)} = -(M^{\mathrm{T}}M)^{-1}(M^{\mathrm{T}}N). \tag{33}$$

5 Simulation study

The purpose of this section is to establish an example to verify the theoretical results. Consider the affine CT nonlinear systems described by [17]

$$\dot{x}_1 = -x_1 + x_2 + 2x_2^3,\ \dot{x}_2 = f(x) + g(x)u, \tag{34}$$

where
$f(x) = \frac{1}{2}(x_1 + x_2) + \frac{1}{2}x_2(1 + 2x_2^2)\sin^2(x_1)$, $g(x) = \sin(x_1)$. The control objective is to regulate the system while minimizing the quadratic functional of the states and the control

$$V(x(t)) = \int_t^{\infty} e^{\alpha(\tau - t)}(Q(x(\tau)) + u^2(x(\tau)))\mathrm{d}\tau,$$

where $Q(x) = x_1^2 + x_2^2 + 2x_2^4$. In order to find the optimal control, we apply polynomials to approximate the cost function as follows:

$$V_8^{(i)}(x_1, x_2) = \omega_1 x_1^2 + \omega_2 x_1 x_2 + \omega_3 x_2^2 + \omega_4 x_1^4 + \omega_5 x_1^3 x_2 + \omega_6 x_1^2 x_2^2 + \omega_7 x_1 x_2^3 + \omega_8 x_2^4.$$

The initial stabilizing control is $u^{(0)}(x) = -\sin(x_1)(1.5x_1 - 0.1x_1^2x_2 + 6x_2^3)$, and the initial state is $x_0 = [1, 1]^{\mathrm{T}}$. The simulation was conducted by using data collected from the system (34) on $\Omega = [-1, 1] \times [-1, 1]$ at $T = 0.25$ s. Now, we provide the convergence of parameters by decreasing the discount factor α. It is significant to note that, though Theorem 2 guarantees the convergence to the optimal control with $\alpha > 0$, the discount factor α cannot be selected as a large number in this example. If α is selected to be a large number, the result of simulation will become rather oscillatory before it comes to convergence. To make matter worse, it might sometimes turn out to be divergent in finite time. Actually, in real control engineering, α cannot be chosen very large. Hence, for convenience, we select $0 < \alpha < 1$ in this example. When $\alpha = 0.4$, by employing the algorithm, the result of simulation is presented by Fig. 1. When $\alpha = 0.2$, the result of simulation is shown by Fig. 2. Though the discount factor is $\alpha \in (0, 1)$, we can let $\alpha \to 0$. If assuming that $\alpha = 0$, we obtain Fig. 3, which is similar with [17]. From Figs. 1, 2 and 3, we know that it costs different time for the parameters of NNs to converge to the coefficients of the value function. When $\alpha = 0.4$, the convergence of parameters needs about 25 s. Meanwhile, there are several oscillations before the parameters become convergent. When $\alpha = 0.2$, the

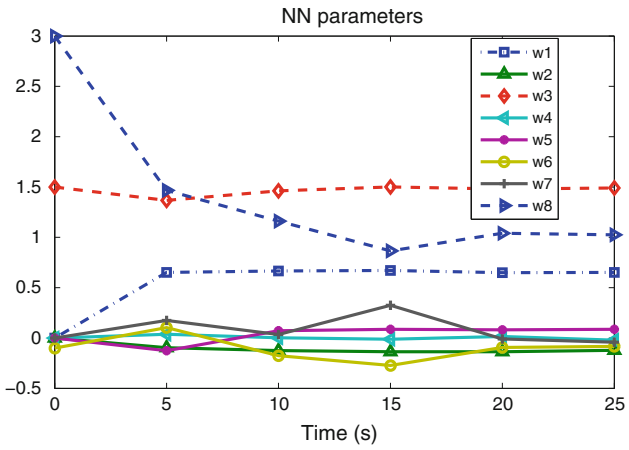

Fig. 1 The convergence of weight parameters, $\alpha = 0.4$

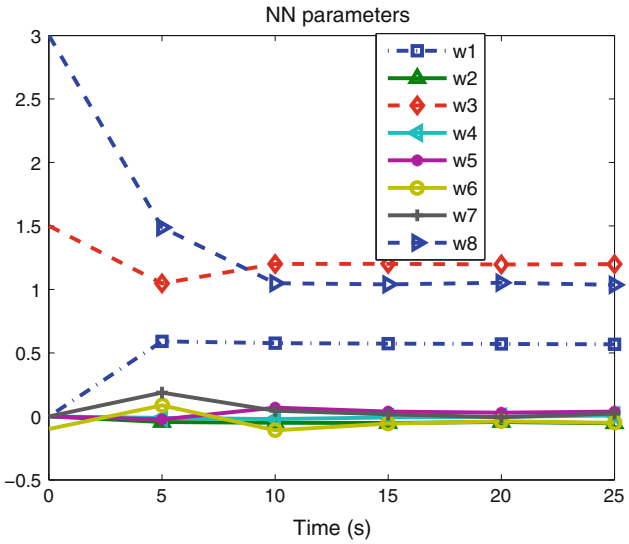

Fig. 2 The convergence of weight parameters, $\alpha = 0.2$

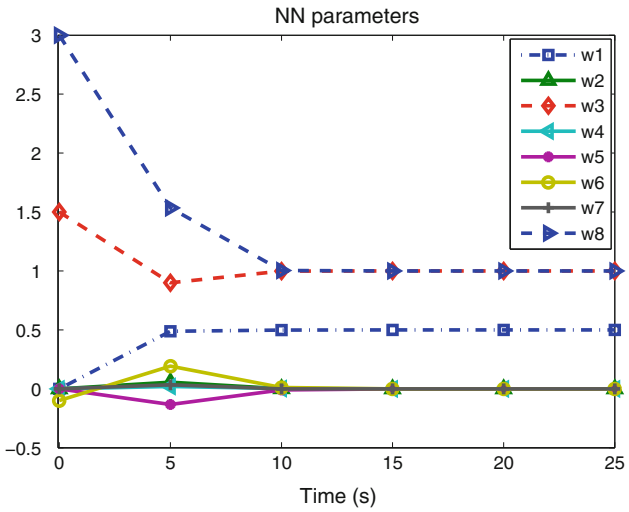

Fig. 3 The convergence of weight parameters, $\alpha = 0$

convergence of parameters needs about 20 s. Moreover, there exist seldom oscillations before the parameters turn to convergence. When $\alpha = 0$, the convergence of parameters needs about 10s. Furthermore, it turns out to be less oscillations than Fig. 2 before the parameters of NNs converging to the coefficients of the value function. From the above analysis, one can come to the conclusion that, when the discount factor is smaller, the rate for the parameters of NNs converging to the coefficients of the value function is higher, and the oscillation is less, and vice versa. Hence, the discount factor α has a significant impact on the performance of the convergence to the optimal cost and the optimal control.

6 Conclusions

In this paper, we investigated the adaptive optimal control problem for infinite horizon cost with a discount factor involved. Without knowledge of the internal dynamics of the nonlinear system and by employing PI, a suboptimal control is obtained. And we find that the impact of the discount factor for CT nonlinear systems is rather different from its influence on DT nonlinear systems. For [14 and 15], there is no significant difference for DT nonlinear systems whether the discount factor is involved in the performance index or not. However, in this paper, not only the HJB equation is modified, but also the convergence to the optimal control has a close connection to the discount factor. Recently, in [24], the discount factor for a DT linear system was discussed, and it presented that the discount factor could significantly alleviate the deleterious effects of probing noise. Whether the discount factor for CT nonlinear systems has the same characteristic or better qualities than DT systems, it is still unknown. These will be investigated in our future work.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grants 61034002, 61233001, and 61273140.

References

1. Pontryagin LS (1959) Optimal control processes. Uspehi Mat Nauk (in Russian) 14:3–20
2. Bellman RE (1957) Dynamic programming. Princeton University Press, New Jersey
3. Lewis FL, Syrmos VL (1995) Optimal control. John Wiley, New York
4. Kailath T (1973) Some new algorithms for recursive estimation in constant linear systems. IEEE Trans Inf Theory 19(6):750–760
5. Laub AJ (1979) A Schur method for solving algebraic Riccati equations. IEEE Trans Autom Control 24(6):913–921
6. Moris K, Navasca C (2006) Iterative solution of algebraic Riccati equations for damped systems. In: Proceedings of 45th IEEE conference on decision and control, San Diego, CA, pp 2436–2440
7. Saridis GN, Lee CS (1979) An approximation theory of optimal control for trainable manipulators. IEEE Trans Syst Man Cybern 9(3):152–159
8. Beard R, Saridis G, Wen J (1997) Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation. Automatica 33(12):2159–2177
9. Abu-Khalaf M, Lewis FL (2005) Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. Automatica 41(5):779–791
10. Murray JJ, Cox CJ, Lendaris GG, Saeks R (2002) Adaptive dynamic programming. IEEE Transa Syst Man Cybern C Appl Rev 32(2):140–153
11. Werbos PJ (1992) Approximate dynamic programming for real-time control and neural modeling. In: White DA, Sofge DA (eds) Handbook of intelligent control: neural, fuzzy, and adaptive approaches. Van Nostrand Reinhold, New York, pp 493–525
12. Wang FY, Zhang H, Liu D (2009) Adaptive dynamic programming: an introduction. IEEE Comput Intell Mag 4(2):39–47
13. Lewis FL, Vrabie D (2009) Reinforcement learning and adaptive dynamic programming for feedback control. IEEE Circuits Syst Mag 9(3):32–50
14. Al-Tamimi A, Lewis FL, Abu-Khalaf M (2008) Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof. IEEE Trans Syst Man Cybern B Cybern 38(4):943–949
15. Liu D, Wang D, Zhao D (2011) Neural-network-based optimal control for a class of nonlinear discrete-time systems with control constraints using the iterative GDHP algorithm. In: Proceedings of international joint conference on neural networks, San Jose, CA, pp 53–60
16. Howard RA (1960) Dynamic programming and Markov processes. MIT Press, Cambridge
17. Vrabie D, Lewis FL (2009) Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. Neural Netw 22(3):237–246
18. Vamvoudakis KG, Lewis FL (2010) Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. Automatica 46(5):878–888
19. Lee JY, Park JB, Choi YH (2010) A novel generalized value iteration scheme for uncertain continuous-time linear systems. In: Proceedings of the 49th IEEE conference on decision and control, Atlanta, GA, pp 4637–4642
20. Guo L, Cheng DZ, Feng DX (2005) Introduction to control theory: from basic concepts to research frontiers. Science Press (in Chinese), Beijing
21. Rudin W (1976) Principles of mathematical analysis, 3rd edn. McGraw-Hill, New York
22. Hornik K, Stinchcombe M, White H (1990) Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Netw 3(5):551–560
23. Finlayson BA (1972) The method of weighted residuals and variational principles. Academic Press, New York
24. Lewis FL, Vamvoudakis KG (2011) Reinforcement learning for partially observable dynamic processes: adaptive dynamic programming using measured output data. IEEE Trans Syst Man Cybern B Cybern 41(1):14–25