2016 American Control Conference (ACC)
Boston Marriott Copley Place
July 6-8, 2016. Boston, MA, USA

Policy Iteration for Optimal Control of Weakly Coupled Nonlinear Systems with Completely Unknown Dynamics

Chao Li[1], Ding Wang[1], Derong Liu[2], and Haibo He[3]

Abstract— **In this paper, an online learning algorithm based on policy iteration is established to solve the optimal control problem for weakly coupled nonlinear continuous-time systems. Using the weak coupling theory, the original problem is transformed into three reduced-order optimal control problems. To obtain the optimal control laws without system dynamics, we construct an online data-based integral policy iteration algorithm which is used to solve the decoupled optimal control problems. The actor-critic technique based on neural networks and the least squares method are used to implement the model-free learning algorithm. A simulation example is given to verify the applicability of the developed algorithm.**

I. Introduction

Many large-scale systems such as transportation systems, electrical networks, power systems, and chemical reactors are naturally weakly coupled [1], [2]. A common challenging problem for these real physical systems is the optimal control. A traditional approach splits the large-scale optimal control problem into some related sub-problems using the decentralized control method [3], [4]. But this approach neglects the coupling effect and the obtained results usually do not have an ideal performance. Since the weakly coupled linear systems were introduced to the control systems community by Kokotovic [5], many theoretical aspects of this problem have been studied. The optimal control law is obtained through a decoupling transformation which leads to solving two independent reduced-order optimal control problems [6], [7]. In a similar way, the optimal control problems for weakly coupled bilinear systems have been solved [8], [9]. Due to the intractable form of the Hamilton-Jacobi-Bellman (HJB) equations that arise in the nonlinear optimal control, obtaining closed-form optimal controllers by directly solving the HJB equations is difficult. By using the reduced-order scheme, the optimal control problems for weakly coupled nonlinear systems have been studied. For instance, Kim and Lim [1] constructed an optimal control law for the weakly coupled nonlinear system based on the solutions of two independent reduced-order HJB equations using successive Galerkin approximation. Carrillo et al. [2] proposed a model-based algorithm for controlling weakly coupled nonlinear systems using the current data and previously stored data with a three-critics/four-actors approximator structure. For large-scale systems, there are many difficulties to obtain the exact knowledge of their dynamics . Therefore, a kind of model-free algorithms is needed to solve the weakly coupled optimal control problem with unknown system dynamics.

Due to the "curse of dimensionality" [10], dynamic programming which provides a method for determining the optimal control laws is often computationally untenable even in the case of completely known dynamics. Adaptive dynamic programming (ADP) and reinforcement learning (RL) relax the need for a exact model of the dynamic systems by using compact parameterized approximators when solving the HJB equations. ADP is an effective computational method due to its optimal learning capabilities [11]–[20]. RL has attracted increasing attention and it can find the optimal policy interactively [21]–[24]. Value iteration and policy iteration (PI) are utilized to solve the HJB equations as main methods of ADP-based and RL-based optimal control. Vrabie and Lewis [25] obtained the direct adaptive optimal control law with partial system dynamics by the established integral RL algorithm for nonlinear continuous-time systems. With completely unknown dynamics, Jiang and Jiang [26] presented a novel PI method to solve the optimal control problem for linear continuous-time systems. Lee et al. [27], [28] derived an integral Q-learning approach for nonlinear system optimal control without the exact knowledge of system dynamics. Li et al. [29] developed an integral RL algorithm to solve two-player zero-sum differential games with completely unknown linear system dynamics. Liu et al. [30] established an online model-free synchronous approximate optimal learning algorithm to solve a multiplayer non-zero-sum differential game.

Among ADP-based and RL-based algorithms, there are few results about the weakly coupled systems. The novelty of this paper is that we establish an online learning algorithm to solve the optimal control for weakly coupled nonlinear systems with completely unknown dynamics. We formulate the original problem into three reduced-order optimal control problems by partitioning the HJB equation. To obtain the optimal control laws without system dynamics, we construct the data-based integral PI algorithm to solve the decoupled optimal control problems. The actor-critic technique based

This work was supported in part by the National Natural Science Foundation of China under Grants 61233001, 61273140, 61304086, 61374105, 61533017, U1501251, and 51529701, in part by Beijing Natural Science Foundation under Grant 4162065, in part by the US National Science Foundation under Grants ECCS 1053717 and IIS-1526835, in part by the Army Research Office under Grant W911NF-12-1-0378, and in part by the Early Career Development Award of SKLMCCS.

[1]C. Li and D. Wang are with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (lichao2012@ia.ac.cn; ding.wang@ia.ac.cn).

[2]D. Liu is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (derong@ustb.edu.cn).

[3]H. He is with the Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881 USA (he@ele.uri.edu).

on neural networks (NNs) and the least squares method are used to implement the integral PI algorithm. We demonstrate the effectiveness of the optimal control law by a simulation example.

The rest of this paper is organized as follows. In Section II, we present the optimal control problem for weakly coupled nonlinear systems. In Section III, we transform the original problem into three reduced-order optimal control problems and establish an online learning algorithm using model-free integral PI with unknown system dynamics. A simulation example is provided to illustrate the effectiveness of the derived optimal control policy in Section IV. We give the conclusion with a few remarks in Section V.

II. Problem Formulation

Consider the weakly coupled nonlinear continuous-time system

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}f_{11}(x_1)+\varepsilon f_{12}(x)\\ \varepsilon f_{21}(x)+f_{22}(x_2)\end{bmatrix} + \begin{bmatrix}g_{11}(x_1) & \varepsilon g_{12}(x)\\ \varepsilon g_{21}(x) & g_{22}(x_2)\end{bmatrix}\begin{bmatrix}u_{11}(t)+\varepsilon u_{12}(t)\\ \varepsilon u_{21}(t)+u_{22}(t)\end{bmatrix}, \quad (1)$$

where $x_1(t)\in\mathbb{R}^{n_1}$, $x_2(t)\in\mathbb{R}^{n_2}$ are the system state vectors, $u_{11}(t),u_{12}(t)\in\mathbb{R}^{m_1}$, $u_{21}(t),u_{22}(t)\in\mathbb{R}^{m_2}$ are the control input vectors, ε is a small positive coupling parameter. Using the following equations

$$x(t)=\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}, f(x)=\begin{bmatrix}f_{11}(x_1)+\varepsilon f_{12}(x)\\ \varepsilon f_{21}(x)+f_{22}(x_2)\end{bmatrix},$$
$$g(x)=\begin{bmatrix}g_{11}(x_1) & \varepsilon g_{12}(x)\\ \varepsilon g_{21}(x) & g_{22}(x_2)\end{bmatrix}, u(t)=\begin{bmatrix}u_{11}(t)+\varepsilon u_{12}(t)\\ \varepsilon u_{21}(t)+u_{22}(t)\end{bmatrix},$$

the system dynamics (1) can be rewritten as

$$\dot{x}(t)=f(x)+g(x)u(t). \quad (2)$$

We assume that $f\colon\mathbb{R}^n\to\mathbb{R}^n$ and $g\colon\mathbb{R}^n\to\mathbb{R}^{n\times m}$ are Lipschitz continuous on the set $\Omega\subseteq\mathbb{R}^n$ and $f(0)=0$, where $n=n_1+n_2$, $m=m_1+m_2$.

The main purpose of the optimal control problem is to find the optimal control law $u^*(x(t))$ to control system (2) with minimum expenditure of control effort. For this reason, the value function are chosen as

$$V(x(t))=\int_t^\infty \left[x^{\mathsf{T}}(\tau)Qx(\tau)+u^{\mathsf{T}}(\tau)Ru(\tau)\right]\mathrm{d}\tau, \quad (3)$$

where $Q\in\mathbb{R}^{n\times n}$ and $R\in\mathbb{R}^{m\times m}$ are positive definite symmetric matrices , and $r(x,u)=x^{\mathsf{T}}(t)Qx(t)+u^{\mathsf{T}}(t)Ru(t)$ is the utility function. The matrices Q and R possess the following weakly coupled structures,

$$Q=\begin{bmatrix}Q_1 & \varepsilon Q_\varepsilon\\ \varepsilon Q_\varepsilon & Q_2\end{bmatrix}, \qquad R=\begin{bmatrix}R_1 & 0\\ 0 & R_2\end{bmatrix}.$$

The optimal value function can be formulated as

$$V^*(x(t))=\min_u\int_t^\infty \left[x^{\mathsf{T}}(\tau)Qx(\tau)+u^{\mathsf{T}}(\tau)Ru(\tau)\right]\mathrm{d}\tau.$$

The Hamiltonian function of system (2) is defined by

$$H(x,u,V_x)=V_x^{\mathsf{T}}[f(x)+g(x)u]+r(x,u), \quad (4)$$

with $V(0)=0$, the term $V_x=\partial V(x)/\partial x$ denotes the partial derivative of the value function with respect to the state. By minimizing the Hamiltonian function (4), the optimal control law can be obtained as

$$u^*(x)=\arg\min_u H(x,u,V_x)=-\frac{1}{2}R^{-1}g^{\mathsf{T}}(x)V_x^*. \quad (5)$$

From the optimal control theory, it is well known that the optimal value function $V^*(x)$ is a unique positive-definite solution of the following HJB equation

$$0=V_x^{*\mathsf{T}}[f(x)+g(x)u^*(x)]+r(x,u^*(x)). \quad (6)$$

III. Computational Controller Design using Online Model-free Policy Iteration Algorithm

In this section, we formulate the original problem into three reduced-order optimal control problems by partitioning the HJB equation. To obtain the optimal control law without system dynamics, we construct the model-free integral PI algorithm to solve the decoupled optimal control problems. To implement this algorithm, the actor-critic technique based on NNs and the least squares method are used.

A. Problem Transformation

According to the reduced-order scheme [1], setting $\varepsilon^2=0$, the value function (3) can be partitioned as

$$V(x(t))=V_1(x_1(t))+V_2(x_2(t))+\varepsilon V_\varepsilon(x(t)),$$

where

$$V_1(x_1(t))=\int_t^\infty\left[x_1^{\mathsf{T}}Q_1x_1+u_{11}^{\mathsf{T}}R_1u_{11}\right]\mathrm{d}\tau,$$
$$V_2(x_2(t))=\int_t^\infty\left[x_2^{\mathsf{T}}Q_2x_2+u_{22}^{\mathsf{T}}R_2u_{22}\right]\mathrm{d}\tau,$$
$$V_\varepsilon(x(t))=2\int_t^\infty\left[x_1^{\mathsf{T}}Q_\varepsilon x_2+u_{11}^{\mathsf{T}}R_1u_{12}+u_{22}^{\mathsf{T}}R_2u_{21}\right]\mathrm{d}\tau.$$

We give the following definitions

$$V_{1x_1}=\frac{\partial V_1}{\partial x_1},\quad V_{2x_2}=\frac{\partial V_2}{\partial x_2},$$
$$V_{\varepsilon x_1}=\frac{\partial V_\varepsilon}{\partial x_1},\quad V_{\varepsilon x_2}=\frac{\partial V_\varepsilon}{\partial x_2}.$$

Partitioning the HJB (6), we get an $\mathcal{O}(\varepsilon^2)$ approximation in terms of three reduced-order decoupled HJB equations

$$\begin{aligned}0=&V_{1x_1}^{*\mathsf{T}}[f_{11}(x_1)+g_{11}(x_1)u_{11}^*(x_1)]\\ &+x_1^{\mathsf{T}}Q_1x_1+u_{11}^{*\mathsf{T}}(x_1)R_1u_{11}^*(x_1),\\ 0=&V_{2x_2}^{*\mathsf{T}}[f_{22}(x_2)+g_{22}(x_2)u_{22}^*(x_2)]\\ &+x_2^{\mathsf{T}}Q_2x_2+u_{22}^{*\mathsf{T}}(x_2)R_2u_{22}^*(x_2),\\ 0=&V_{1x_1}^{*\mathsf{T}}f_{12}(x)+V_{2x_2}^{*\mathsf{T}}f_{21}(x)+V_{\varepsilon x_1}^{*\mathsf{T}}f_{11}(x_1)\\ &+V_{\varepsilon x_2}^{*\mathsf{T}}f_{22}(x_2)+2x_1^{\mathsf{T}}Q_\varepsilon x_2\\ &-2u_{11}^{*\mathsf{T}}(x_1)R_1u_{12}^*(x)-2u_{22}^{*\mathsf{T}}(x_2)R_2u_{21}^*(x).\end{aligned}$$

The optimal control law (5) can be partitioned as

$$\begin{aligned}
u_{11}^*(x_1) &= -\frac{1}{2}R_1^{-1}g_{11}^{\mathsf{T}}(x_1)V_{1x_1}^*,\\
u_{12}^*(x) &= -\frac{1}{2}R_1^{-1}[g_{11}^{\mathsf{T}}(x_1)V_{\varepsilon x_1}^* + g_{21}^{\mathsf{T}}(x)V_{2x_2}^*],\\
u_{21}^*(x) &= -\frac{1}{2}R_2^{-1}[g_{22}^{\mathsf{T}}(x_2)V_{\varepsilon x_2}^* + g_{12}^{\mathsf{T}}(x)V_{1x_1}^*],\\
u_{22}^*(x_2) &= -\frac{1}{2}R_2^{-1}g_{22}^{\mathsf{T}}(x_2)V_{2x_2}^*.
\end{aligned}$$

According to the optimal control theory, $u_{11}^*(x_1)$ is the optimal control law for the subsystem 1

$$\dot{x}_1(t) = f_{11}(x_1) + g_{11}(x_1)u_{11}(t),$$

$u_{22}^*(x_2)$ is the optimal control law for the subsystem 2

$$\dot{x}_2(t) = f_{22}(x_2) + g_{22}(x_2)u_{22}(t),$$

$u_{12}^*(x)$ and $u_{21}^*(x)$ can be solved from the following integral equation related to the subsystem 3

$$V_3^*(x) = 2\int_t^\infty \big[u_{11}^{*\mathsf{T}}(x_1)R_1u_{12}^*(x) + u_{22}^{*\mathsf{T}}(x_2)R_2u_{21}^*(x) - x_1^{\mathsf{T}}Q_\varepsilon x_2\big]\mathrm{d}\tau,$$

where $V_3^*(x) = V_\varepsilon^*(x) - 4\int_t^\infty x_1^{\mathsf{T}}Q_\varepsilon x_2\mathrm{d}\tau$ with $V_3^*(0) = 0$.

B. Model-free Algorithm

To deal with the optimal control problem with completely unknown dynamics, we develop a data-based online integral PI algorithm. Consider the following nonlinear system explored by a known bounded piecewise continuous probing signal $e(t)$

$$\dot{x}(t) = f(x) + g(x)[u(t) + e(t)],$$

where

$$u(t) + e(t) = \begin{bmatrix}[u_{11}(t) + e_1(t)] + \varepsilon u_{12}(t)\\ \varepsilon u_{21}(t) + [u_{22}(t) + e_2(t)]\end{bmatrix}.$$

Now we consider the subsystem 1 with exploration signal

$$\dot{x}_1(t) = f_{11}(x_1) + g_{11}(x_1)[u_{11}(t) + e_1(t)]. \tag{7}$$

The derivative of the value function $V_1(x_1(t))$ with respect to time along the trajectory of the explored subsystem (7) can be calculated as

$$\begin{aligned}
\dot{V}_1(x_1(t)) &= V_{1x_1}^{\mathsf{T}}\Big[f_{11}(x_1) + g_{11}(x_1)[u_{11}(t) + e_1(t)]\Big]\\
&= -r_1(x_1, u_{11}(x_1)) + V_{1x_1}^{\mathsf{T}}g_{11}(x_1)e_1(t),
\end{aligned} \tag{8}$$

where $r_1(x_1, u_{11}(x_1)) = x_1^{\mathsf{T}}Q_1x_1 + u_{11}^{\mathsf{T}}(x_1)R_1u_{11}(x_1)$ is the utility function for the subsystem 1. Based on the traditional PI algorithm and using the representations $V_1^i(x_1(t))$ and $u_{11}^i(x_1)$, the policy improvement can be represented as

$$u_{11}^{i+1}(x_1) = -\frac{1}{2}R_1^{-1}g_{11}^{\mathsf{T}}(x_1)V_{1x_1}^i, \tag{9}$$

where i is the iterative index. Integrating (8) form t to $t+T$ and considering the policy improvement (9), we have

$$\begin{aligned}
V_1^i(x_1(t)) - V_1^i(x_1(t+T)) &= \int_t^{t+T} r_1(x_1, u_{11}^i(x_1)\mathrm{d}\tau\\
&\quad + 2\int_t^{t+T}(u_{11}^{i+1}(x_1))^{\mathsf{T}}R_1e_1(\tau)\mathrm{d}\tau,
\end{aligned} \tag{10}$$

where the time interval $T > 0$. Since $f_{11}(x_1)$ and $g_{11}(x_1)$ do not appear in the integral equation (10), the integral PI algorithm can be done without the complete system dynamics. Thus, we describe the online model-free integral PI algorithm in Algorithm 1.

Algorithm 1 (Integral Policy Iteration Algorithm)

1: Give a small positive real number ϵ. Let $i = 0$ and start with an initial admissible control law $u_{11}^0(x_1)$.
2: Policy Evaluation and Improvement: Based on the control policy $u_{11}^i(x_1)$, solve $V_1^i(x_1)$ and $u_{11}^{i+1}(x_1)$ from the integral equation (10).
3: If $\|u_{11}^{i+1}(x_1) - u_{11}^i(x_1)\| \le \epsilon$, stop and obtain the approximate optimal control law for the subsystem 1; else, set $i = i + 1$ and go to Step 2.

Theorem 1: Give an initial admissible control law $u_{11}^0(x_1)$ for the subsystem 1. Using the integral PI algorithm established in Algorithm 1, the value function and the control law converge to the optimal ones as $i \to \infty$, i.e.,

$$V_1^i(x_1) \to V_1^*(x_1), \quad u_{11}^i(x_1) \to u_{11}^*(x_1).$$

Proof: If the initial control law $u_{11}^0(x_1)$ is admissible, during the iteration process of Algorithm 1, all the subsequent control laws will be admissible [25]. Considering the formation process of (10) and the equivalence between (8) and the traditional PI algorithm, the iteration result in Algorithm 1 will converge to the solution of the HJB equation. So we can conclude that the proposed integral PI algorithm will converge to the solution of the optimal control problem for the subsystem 1 without using the knowledge of system dynamics. The proof is completed. ■

Subsystem 2 has the same structure with subsystem 1, we can use Algorithm 1 to calculate the optimal control law $u_{22}^*(x_2)$ applying some replacements. After obtaining the optimal control laws $u_{11}^*(x_1)$ and $u_{22}^*(x_2)$, we derive the following equation to solve for $u_{12}^*(x)$ and $u_{21}^*(x)$ iteratively,

$$\begin{aligned}
&V_3^i(x(t)) - V_3^i(x(t+T)) = -2\int_t^{t+T} x_1^{\mathsf{T}}Q_\varepsilon x_2\mathrm{d}\tau\\
&+ 2\int_t^{t+T}[u_{11}^{*\mathsf{T}}(x_1)R_1u_{12}^i(x) + u_{22}^{*\mathsf{T}}(x_2)R_2u_{21}^i(x)]\mathrm{d}\tau\\
&+ 2\int_t^{t+T}[(u_{12}^{i+1}(x))^{\mathsf{T}}R_1e_1(\tau) + (u_{21}^{i+1}(x))^{\mathsf{T}}R_2e_2(\tau)]\mathrm{d}\tau.
\end{aligned}$$

Using this equation to replace (10) in Algorithm 1, we can obtain $u_{12}^*(x)$ and $u_{21}^*(x)$.

C. Algorithm Implementation

For the subsystem 1, we assume that $V_1^i(x_1)$ and $u_{11}^{i+1}(x_1)$ can be represented on a compact set Ω by single-layer networks as

$$V_1^i(x_1) = \sum_{j=1}^{N_c} \omega_{ij}\phi_j(x_1) + \delta_c(x_1),$$
$$u_{11,p}^{i+1}(x_1) = \sum_{j=1}^{N_a} \nu_{ij,p}\psi_j(x_1) + \delta_{a,p}(x_1),$$

where $p = 1, 2, \ldots, m_1$, $\omega_{ij} \in \mathbb{R}$ and $\nu_{ij,p} \in \mathbb{R}$ are the bounded ideal weight parameters which are unknown and will be calculated by the established integral PI algorithm, $\phi_j(x_1) \in \mathbb{R}$ and $\psi_j(x_1) \in \mathbb{R}$, $\{\phi_j\}_{j=1}^{N_c}$ and $\{\psi_j\}_{j=1}^{N_a}$ are the sequences of real-valued activation functions that are linearly independent and complete, and $\delta_c(x) \in \mathbb{R}$ and $\delta_{a,p}(x) \in \mathbb{R}$ are the bounded NN approximation errors. Since ω_{ij} and $\nu_{ij,p}$ are unknown, the outputs of the critic NN and the action NN are represented as

$$\hat{V}_1^i(x_1) = \sum_{j=1}^{N_c} \hat{\omega}_{ij}\phi_j(x_1) = \hat{\omega}_i^{\mathsf{T}}\phi(x_1), \tag{11}$$
$$\hat{u}_{11,p}^{i+1}(x_1) = \sum_{j=1}^{N_a} \hat{\nu}_{ij,p}\psi_j(x_1) = \hat{\nu}_{i,p}^{\mathsf{T}}\psi(x_1), \tag{12}$$

where $\hat{\omega}_i$ and $\hat{\nu}_{i,p}$ are the current estimated weights, and

$$\begin{aligned}
\phi(x_1) &= [\phi_1(x_1), \phi_2(x_1), \ldots, \phi_{N_c}(x_1)]^{\mathsf{T}} \in \mathbb{R}^{N_c},\\
\psi(x_1) &= [\psi_1(x_1), \psi_2(x_1), \ldots, \psi_{N_a}(x_1)]^{\mathsf{T}} \in \mathbb{R}^{N_a},\\
\hat{\omega}_i &= [\hat{\omega}_{i1}, \hat{\omega}_{i2}, \ldots, \hat{\omega}_{iN_c}]^{\mathsf{T}} \in \mathbb{R}^{N_c},\\
\hat{\nu}_{i,p} &= [\hat{\nu}_{i1,p}, \hat{\nu}_{i2,p}, \ldots, \hat{\nu}_{iN_a,p}]^{\mathsf{T}} \in \mathbb{R}^{N_a},\\
\hat{\nu}_i^{\mathsf{T}} &= [\hat{\nu}_{i,1}, \hat{\nu}_{i,2}, \ldots, \hat{\nu}_{i,m_1}]^{\mathsf{T}} \in \mathbb{R}^{m_1 \times N_a}.
\end{aligned}$$

Define $\mathrm{col}\{\hat{\nu}_i^{\mathsf{T}}\} = [\hat{\nu}_{i,1}^{\mathsf{T}}, \hat{\nu}_{i,2}^{\mathsf{T}}, \ldots, \hat{\nu}_{i,m_1}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{m_1 N_a}$, then

$$\begin{aligned}
(\hat{u}_{11}^{i+1}(x_1))^{\mathsf{T}} R_1 e_1(t) &= (\hat{\nu}_i^{\mathsf{T}}\psi(x_1))^{\mathsf{T}} R_1 e_1(t)\\
&= (\psi(x_1) \otimes (R_1 e_1(t)))^{\mathsf{T}} \mathrm{col}\{\hat{\nu}_i^{\mathsf{T}}\},
\end{aligned}$$

where $\otimes$ represents the Kronecker product. Substituting the expressions (11) and (12) into the integral equation (10), we obtain the following general form

$$\lambda_k^{\mathsf{T}} \begin{bmatrix} \hat{\omega}_i \\ \mathrm{col}\{\hat{\nu}_i^{\mathsf{T}}\} \end{bmatrix} = \theta_k \tag{13}$$

with

$$\begin{aligned}
\theta_k &= \int_{t+(k-1)T}^{t+kT} [x_1^{\mathsf{T}} Q_1 x_1 + \hat{u}_{11}^{i\mathsf{T}}(x_1) R_1 \hat{u}_{11}^i(x_1)]\mathrm{d}\tau.\\
\lambda_k &= \Bigg[\Big(\phi(x_1(t+(k-1)T)) - \phi(x_1(t+kT))\Big)^{\mathsf{T}},\\
&\quad -2\int_{t+(k-1)T}^{t+kT} (\psi(x_1) \otimes (R_1 e_1(\tau)))^{\mathsf{T}} \mathrm{d}\tau\Bigg]^{\mathsf{T}},
\end{aligned}$$

where the data collection time is from $t+(k-1)T$ to $t+kT$. We cannot guarantee the uniqueness of the solution of (13) which is only a 1-dimensional equation. We use the least squares method to solve the weights over the compact set Ω as in [27]. For any positive integer K, we denote $\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_K]$ and $\Theta = [\theta_1, \theta_2, \ldots, \theta_K]^{\mathsf{T}}$. Then, we have the following K-dimensional equation

$$\Lambda^{\mathsf{T}} \begin{bmatrix} \hat{\omega}_i \\ \mathrm{col}\{\hat{\nu}_i^{\mathsf{T}}\} \end{bmatrix} = \Theta.$$

If Λ^{T} has full column rank, the parameters can be solved by

$$\begin{bmatrix} \hat{\omega}_i \\ \mathrm{col}\{\hat{\nu}_i^{\mathsf{T}}\} \end{bmatrix} = (\Lambda\Lambda^{\mathsf{T}})^{-1}\Lambda\Theta. \tag{14}$$

Therefore, the number of collected points K should be satisfied $K \geq \mathrm{rank}(\Lambda) = N_c + m_1 N_a$, which will make $(\Lambda\Lambda^{\mathsf{T}})^{-1}$ exist. The least squares problem in (14) can be solved in real time by collecting enough data points generated from the explored system (7). The implementation procedures for subsystems 2 and 3 are same as subsystem 1.

IV. Numerical Example

We provide a simulation example to demonstrate the effectiveness of the optimal control law established for the weakly coupled system.

We consider a weakly coupled nonlinear system (1) with the following parameters

$$\begin{aligned}
f_{11}(x_1) &= \begin{bmatrix} -1.93x_{11}^2 \\ -1.394x_{11}x_{12} \end{bmatrix},\\
f_{12}(x) &= \begin{bmatrix} 0 \\ -4.26x_{21}x_{22} \end{bmatrix},\\
f_{21}(x) &= \begin{bmatrix} -1.3x_{12}^2 \\ 0.95x_{11}x_{21} - 1.03x_{12}x_{22} \end{bmatrix},\\
f_{22}(x_2) &= \begin{bmatrix} -0.63x_{21}^2 \\ 0.413x_{21} - 0.426x_{22} \end{bmatrix},\\
g_{11}(x_1) &= \begin{bmatrix} -1.274x_{11}^2 \\ 0 \end{bmatrix}, \qquad g_{12}(x) = \begin{bmatrix} 0 \\ -6.5x_{22} \end{bmatrix},\\
g_{21}(x) &= \begin{bmatrix} 0.75x_{11} \\ 0 \end{bmatrix}, \qquad g_{22}(x_2) = \begin{bmatrix} -0.718x_{21} \\ 0 \end{bmatrix}.
\end{aligned}$$

$x_1 = [x_{11}, x_{12}]^{\mathsf{T}} \in \mathbb{R}^2$ and $u_{11}(x_1) \in \mathbb{R}$ are the state and control variables of subsystem 1, and $x_2 = [x_{21}, x_{22}]^{\mathsf{T}} \in \mathbb{R}^2$ and $u_{22}(x_2) \in \mathbb{R}$ are the state and control variables of subsystem 2. The initial state is $x(0) = [3.4, 2.7, 4.3, 1.2]^{\mathsf{T}}$. The weak coupling parameter $\varepsilon = 0.05$. The matrices Q and R are chosen as

$$R = Q_1 = Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad Q_\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & 0.05 \end{bmatrix}.$$

During the simulation, the exact knowledge of the dynamics is assumed to be completely unknown. We use the established algorithm to solve the optimal control problem.

For the subsystem 1

$$\dot{x}_1 = \begin{bmatrix} -1.93x_{11}^2 \\ -1.394x_{11}x_{12} \end{bmatrix} + \begin{bmatrix} -1.274x_{11}^2 \\ 0 \end{bmatrix} u_{11}(x_1),$$

the weight vectors of the critic and action networks are represented as

$$\begin{aligned}
\hat{\omega}^1 &= [\hat{\omega}_1^1, \hat{\omega}_2^1, \hat{\omega}_3^1]^{\mathsf{T}},\\
\hat{\nu}^1 &= [\hat{\nu}_1^1, \hat{\nu}_2^1]^{\mathsf{T}}.
\end{aligned}$$

We choose the activation functions as

$$\phi^1(x_1) = [x_{11}^2, x_{11}x_{12}, x_{12}^2]^\mathsf{T},$$
$$\psi^1(x_1) = [x_{11}x_{12}, x_{12}^2]^\mathsf{T}.$$

According to the activation functions, we know $N_c^1 = 3$ and $N_a^1 = 2$, so the simulation can be conducted with $K^1 = 10$. The initial weights are set as $\hat{\omega}^1 = [0, 0, 0]^\mathsf{T}$ and $\hat{\nu}^1 = [2, 1]^\mathsf{T}$. During the learning process, the period time $T = 0.1s$ and the probing signal $e_1(t) = 3\sin(2\pi t) + 3\cos(2\pi t)$ are used. After 10 samples are obtained, the least squares problem can be solved. Thus the weights of the networks are updated every $1s$. Fig. 1 illustrates the weights evolutions of the action network 1. The precision $\epsilon = 10^{-4}$ is achieved after 52 iterations. At time $t = 52s$, $\hat{\nu}^{1*} = [-1.2557, -0.1067]^\mathsf{T}$.

We choose the activation functions for the subsystem 2 as

$$\phi^2(x_2) = [x_{21}^2, x_{21}x_{22}, x_{22}^2]^\mathsf{T},$$
$$\psi^2(x_2) = [x_{21}x_{22}, x_{22}^2]^\mathsf{T}.$$

As $N_c^2 = 3$ and $N_a^2 = 2$, the simulation can be conducted with $K^2 = 10$. The initial weights are set as $\hat{\omega}^2 = [0, 0, 0]^\mathsf{T}$ and $\hat{\nu}^2 = [10, 2]^\mathsf{T}$. During the learning process, the period time $T = 0.1s$ and the probing signal $e_2(t) = 5\sin(2\pi t) + 5\cos(2\pi t)$ are used. Fig. 2 illustrates the weights evolutions of the action network 2. The precision $\epsilon = 10^{-4}$ is achieved after 50 iterations. At time $t = 50s$, $\hat{\nu}^{2*} = [-9.9814, 0.0367]^\mathsf{T}$.

For the subsystem 3, the weight vectors of the critic and action networks are represented as

$$\hat{\omega}^3 = [\hat{\omega}_1^3, \hat{\omega}_2^3, \hat{\omega}_3^3, \hat{\omega}_4^3, \hat{\omega}_5^3, \hat{\omega}_6^3]^\mathsf{T},$$
$$\hat{\nu}^3 = [\hat{\nu}_1^3, \hat{\nu}_2^3, \hat{\nu}_3^3, \hat{\nu}_4^3]^\mathsf{T}.$$

We choose the activation functions as

$$\phi^3(x) = [x_{11}^2, x_{11}x_{12}, x_{12}^2, x_{21}^2, x_{21}x_{22}, x_{22}^2]^\mathsf{T},$$
$$\psi^3(x) = [x_{11}x_{12}, x_{12}^2, x_{21}x_{22}, x_{22}^2]^\mathsf{T}.$$

According to the activation functions, we know $N_c^3 = 6$ and $N_a^3 = 4$, so the simulation can be conducted with $K^3 = 10$. The initial weights are set as $\hat{\omega}^3 = [0, 0, 0, 0, 0, 0]^\mathsf{T}$ and $\hat{\nu}^3 = [0, -2, 2, 3]^\mathsf{T}$. During the learning process, the period time $T = 0.1s$ and the probing signals $e_1(t)$, $e_2(t)$ are used. After 10 samples are obtained, the least squares problem can be solved. Thus the weights of the networks are updated every $1s$. Fig. 3 illustrates the weights evolutions of the action network 3. The precision $\epsilon = 10^{-4}$ is achieved after 20 iterations. At time $t = 20s$, $\hat{\nu}^{3*} = [0.3830, 0.0533, -0.0899, -0.9548]^\mathsf{T}$.

According to Section III, we obtain the following optimal control law of the weakly coupled system

$$u^*(x) = \begin{bmatrix} u_{11}^*(x_1) + \varepsilon u_{12}^*(x) \\ \varepsilon u_{21}^*(x) + u_{22}^*(x_2) \end{bmatrix}.$$

We use the optimal control law $u^*(x)$ to control the weakly coupled system for $20s$. The evolution process of the system state and control trajectories shown in Figs. 4 and 5. These simulation results can verify the effectiveness of the integral PI algorithm.

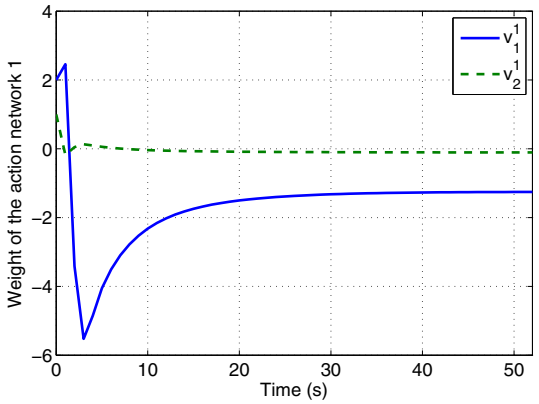

Fig. 1. Weights evolutions of the action network 1.

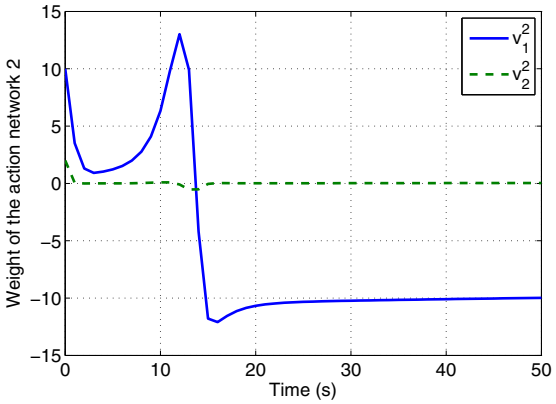

Fig. 2. Weights evolutions of the action network 2.

V. Conclusions

In this paper, a model-free integral PI algorithm for weakly coupled nonlinear systems is developed. The optimal control law is derived using the optimal controllers of the reduced-order subsystems. To solve the reduced-order HJB equations related to the subsystems, we use establish a model-free integral PI algorithm. The actor-critic technique and the least squares method are used to implement the constructed algorithm. The applicability of the developed optimal control law is testified by a simulation example.

References

[1] Y. J. Kim and M. T. Lim, "Parallel optimal control for weakly coupled nonlinear systems using successive Galerkin approximation," *IEEE Transactions on Automatic Control*, vol. 53, no. 6, pp. 1542–1547, July 2008.

[2] L. R. Garcia Carrillo, K. G. Vamvoudakis, and J. P. Hespanha, "Approximate optimal adaptive control for weakly coupled nonlinear systems: A neuro-inspired approach," to appear in International Journal of Adaptive Control and Signal Processing, 2015.

[3] A. Saberi, "On optimality of decentralized control for a class of nonlinear interconnected systems," *Automatica*, vol. 24, no. 1, pp. 101–104, Jan. 1988.

[4] D. Liu, D. Wang, and H. Li, "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*, vol. 25, no. 2, pp. 418–428, Feb. 2014.

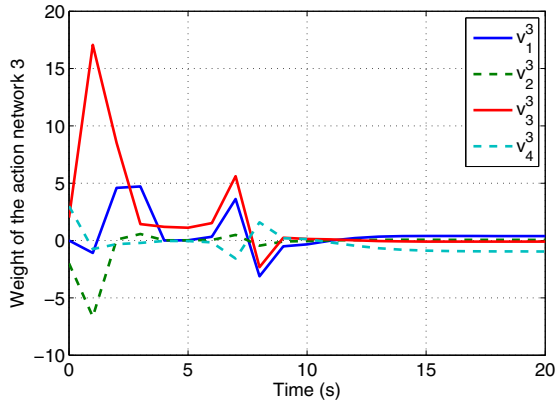

Fig. 3. Weights evolutions of the action network 3.

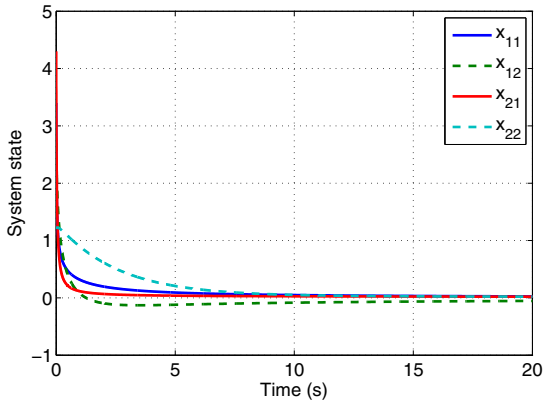

Fig. 4. State trajectory of the weakly coupled system under the developed control law.

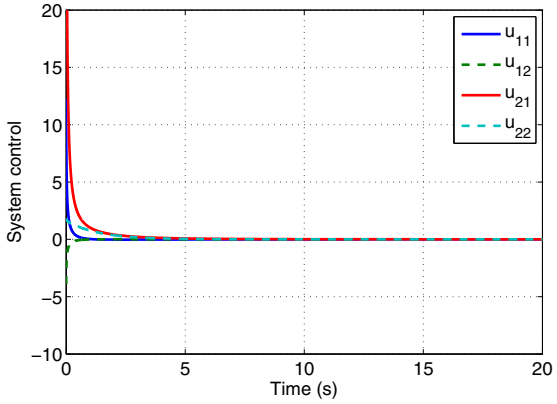

Fig. 5. Control trajectory of the weakly coupled system under the developed control law.

[5] P. Kokotovic, W. Perkins, J. Cruz, and G. DAns, "ϵ-coupling method for near-optimum design of large-scale linear systems," *Proceedings of the Institution of Electrical Engineers*, vol. 116, pp. 889–892, 1969.

[6] Z. Gajic and X. Shen, "Decoupling transformation for weakly coupled linear systems," *International Journal of Control*, vol. 50, pp. 1515–1521, 1989.

[7] Z. Gajic and X. Shen, Parallel Algorithms for Optimal Control of Large Scale Linear Systems. London, U.K.: Springer, 1992.

[8] Z. Aganovic and Z. Gajic, "Optimal control of weakly coupled bilinear systems," *Automatica*, vol. 29, pp. 1591–1593, 1993.

[9] Z. Aganovic and Z. Gajic, Linear Optimal Control of Bilinear Systems: with Applications to Singular Perturbations and Weak Coupling. London, U.K.: Springer, 1995.

[10] R. Bellman, Dynamic Programming. Princeton University Press, 1957.

[11] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 39–47, May 2009.

[12] D. Liu, H. Javaherian, O. Kovalenko, and T. Huang, "Adaptive critic learning techniques for engine torque and air-fuel ratio control," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 38, no. 4, pp. 988–993, Aug. 2008.

[13] Z. Ni, H. He, X. Zhong, and D. V. Prokhorov, "Model-free dual heuristic dynamic programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 8, pp. 1834–1839, Aug. 2015.

[14] Z. Ni, H. He, D. Zhao, X. Xu, and D. V. Prokhorov, "GrDHP: A general utility function representation for dual heuristic dynamic programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 3, pp. 614–627, Mar. 2015.

[15] D. Wang, D. Liu, H. Li, and H. Ma, "Neural-network-based robust optimal control design for a class of uncertain nonlinear systems via adaptive dynamic programming," *Information Sciences*, vol. 282, pp. 167–179, Oct. 2014.

[16] D. Wang, D. Liu, and H. Li, "Policy iteration algorithm for online design of robust control of a class of continuous-time nonlinear systems," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 627–632, Apr. 2014.

[17] X. Yang, D. Liu, and D. Wang, "Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints," *International Journal of Control*, vol. 87, no. 3, pp. 553–566, Mar. 2014.

[18] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming with an application to power systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 7, pp. 1150–1156, July 2013.

[19] C. Li, D. Liu, H. Li, D. Wang, H. Ma, "Neural-Network-Based Decentralized Control of Continuous-Time Nonlinear Interconnected Systems with Unknown Dynamics," *Neurocomputing*, vol. 165, pp. 90–98, Oct. 2015.

[20] H. Li and D. Liu, "Optimal control for discrete-time affine nonlinear systems using general value iteration," *IET Control Theory & Applications*, vol. 6, no. 18, pp. 2725–2736, Dec. 2012.

[21] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, Mar. 2009.

[22] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction, volume 1. Cambridge Univ Press, 1998.

[23] J. Si, A. G. Barto, W. B. Powell, D. C. Wunsch, Handbook of learning and approximate dynamic programming. Los Alamitos: IEEE Press , 2004.

[24] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proceedings of American Control Conference*, volume 3, Baltimore, MD, Jun. 1994, pp. 3475–3479.

[25] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, Mar. 2009.

[26] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, Oct. 2012.

[27] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica*, vol. 48, no. 11, pp. 2850–2859, Nov. 2012.

[28] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning with explorations for continuous-time nonlinear systems," in *Proceedings of The 2012 International Joint Conference on Neural Networks*, Brisbane, QLD, June 2012, pp. 1042–1047.

[29] H. Li, D. Liu, D. Wang, and X. Yang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 3, pp. 706–714, July 2014.

[30] D. Liu, H. Li, and D. Wang, "Online synchronous approximate optimal learning algorithm for multi-player non-zero-sum games with unknown dynamics," *IEEE Transactions on Systems, Man and Cybernetics: Systems*, vol. 44, no. 8, pp. 1015–1027, Aug. 2014.