Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

A set-based model-free reinforcement learning design technique for nonlinear systems

Martin Guay * **Khalid Tourkey Atta** **

* *Department of Chemical Engineering, Queen's University, Kingston, ON, Canada E-mail address:* martin.guay@chee.queensu.ca
** *Control Engineering Group, Lulea University of Technology, 971 87 Lulea, Sweden, E-mail address:*khalid.atta@Ltu.se

Abstract: In this study, we propose an extremum-seeking approach for the approximation of optimal control problems for unknown nonlinear dynamical systems. The technique combines a phasor extremum seeking controller with an reinforcement learning strategy. The learning approach is used to estimate the value function of an optimal control problem of interest. The phasor extremum seeking controller implements the approximate optimal controller. The approach is shown to provide reasonable approximations of optimal control problems without the need for a parameterization of the nonlinear control system. A simulation example are provided to demonstrate the effectiveness of the technique.

Keywords: Extremum-seeking control, Reinforcement learning, phasor approximation

1. INTRODUCTION

Recent developments in learning techniques have inspired control theorists and researchers to integrate control methodologies for the design of data-based control systems. One such technique is reinforcement learning (see Sutton and Barto (1998), Watkins and Dayan (1992) and references therein). In reinforcement learning, a control system is made to adjust its actions according to a meaningful user-defined optimal control problem. The learning process is generally iterative and involves a certain parameterization of the unknown value function (i.e., a solution of the Bellman equation) and, possibly, the optimal controller (Barto et al. (1983), Bertsekas and Tsitsiklis (1995), Busoniu et al. (2010) ,Mehta and Meyn (2009), Sutton et al. (1992), Bradtke et al. (1994)) . The identification of the parameters requires some probing of the system dynamics using some external excitation signal. One common iterative learning approach is Q-learning Watkins and Dayan (1992). In this technique, the value function is estimated by iteratively computing a Q-function. This Q-function can be viewed as a Bellman function that is subject to an existing (sub-optimal) state-feedback controller. At each step, the Q-function yields an estimate of the value function that can be used to define a new state-feedback controller. Convergence is achieved when the Bellman error reaches a suitable tolerance level.

Model-free and model-based approaches have been proposed. It is generally recognized that model-based strategies (such as Vamvoudakis and Lewis (2010)), which requires knowledge of the system's dynamics, provide a more effective learning strategy that results in a fewer number of learning steps. Model-free techniques are increasingly popular in the context of machine learning (such as Mehta and Meyn (2009), Sutton et al. (1992), Bradtke et al. (1994),Bhasin et al. (2013)). Their design is obviously more challenging. For control affine systems, researchers have proposed affine parameterizations of the unknown dynamics in addition to the parameterization requirements of Q-learning. Some techniques have cleverly integrated the unknown dynamics into a single actor-critic approach.

In this study, we propose a novel on-policy model-free reinforcement learning technique. The proposed technique introduces two new element in the solution of Q-learning problems. The first element is a set-based parameter estimation technique that is suitable for nonlinearly parameterized dynamical systems. The technique, originally presented in Adetola et al. (2014), provides an effective mechanism that avoids the actor-critic methodology for Q-learning by providing a unique parameterization. The second element is a phasor extremum-seeking control approach as initially proposed in Atta et al. (2015). In the context of this learning approach, the phasor extremum seeking approach allows one to deal with systems with unknown dynamics. The combination of the two techniques is shown to provide a learning approach that avoids the dual parameterization of the actor-critic approaches and the parameterization of the unknown nonlinear dynamics. The total number of unknown parameters are therefore reduced which improves the convergence properties of the technique and minimizes the extent of external excitation necessary for the estimation of the parameters.

The paper is organized as follows. A problem description of the learning problem along with the key assumptions

are given in Section 2. The application of the set-based nonlinear estimation approach for Q-learning is present in Section 3. The phasor estimation approach is presented in Section 4. The complete integrated model-free mechanism is presented in Section 5 along with the analysis of stability. A simulation example is presented in Section 6 followed by brief conclusions and proposed future work in Section 7.

2. PROBLEM DEFINITION

We consider control affine nonlinear systems described by:

$$\dot{x} = f(x) + g(x)u \tag{1}$$

where $x \in \mathbb{R}^n$ is the vector of state variables, $u \in \mathbb{R}^p$ is the vector of input variables, $f : \mathbb{R}^n \to \mathbb{R}^n$ is a vector valued smooth function of the state variables and $g : \mathbb{R}^n \to \mathbb{R}^{n\times p}$ is a matrix valued function of the state variables.

It is assumed that the state variables are available for measurement but that the dynamics of the system (described by $f(x)$ and $g(x)$) are unknown.

The objective is the compute the control law, $u(t) = \alpha(x(t))$, that minimizes the cost functional:

$$J(x_0, u(t)) = \int_0^\infty Q(x(t)) + u(t)^T R u(t) dt. \tag{2}$$

The minimization of $J(x_0, u(t)$ subject to $u(t)$ has a value function given by:

$$V^*(x) = \inf_{u(t)} \int_0^\infty Q(x(t)) + u(t)^T R u(t) dt. \tag{3}$$

We define the Lie derivative of a function, $V(x)$, along the vector field $f(x)$ and $g(x)$ as:

$$L_f V^* = (\nabla_x V^*) f(x),\ L_g V^* = (\nabla_x V^*) g(x), \tag{4}$$

respectively.

Assuming that $V^*(x)$ is finite valued and continuous differentiable, it satisfies the nonlinear partial differential equation (PDE):

$$\min_u \left(L_f V^* + L_g V^* u + Q(x) + u^T R u \right) = 0. \tag{5}$$

The solution of this PDE leads to the well-known optimal state-feedback controller:

$$\text{argmin}_u \mathcal{H}(x, \nabla_x V^*, u) = -\frac{1}{2} R^{-1} L_g V^{*^T}. \tag{6}$$

We consider term on the left hand side of (5) for some arbitrary input u. This defines the function, often called Q–function, as follows:

$$\mathcal{H}^*(x, u) = L_f V^* + L_g V^* u + Q(x) + u^T R u. \tag{7}$$

The problem considered here is to approximate the optimal controller corresponding to the value function $V^*(x)$ using only the measurement of the state variables. Inspired by existing Q–learning and reinforcement learning techniques, we propose an adaptive approach to the computation of optimal controllers. The approach utilizes a set-based hybrid learning recursive least squares technique described in the next section.

3. SET-BASED LEAST-SQUARES Q–LEARNING

In this section, we propose an alternative estimation algorithm that prevents the need for the standard iterative or actor–critic Q–learning methodologies used to identify the unknown value function and the corresponding optimal state-feedback.

Let us recall a standard actor–critic Q–learning algorithm. The strategy is to consider a functional approximation of the value function of the form:

$$V(x) = W_c^T \phi(x) + e_c(x) \tag{8}$$

where $W_c \in \mathbb{R}^N$ is a vector of unknown constants to be estimated, $\phi : \mathbb{R}^n \to \mathbb{R}^N$ is a vector of smooth basis functions and $e_c(x)$ is the truncation error. The basis functions are such that $\lim_{N\to\infty} \|e_c(x)\| = 0$.

Using this parameterization of the value function, the Q–function can be approximated as follows:

$$\mathcal{H}_c(W_c, \phi, x, u) = W_c^T \frac{\partial \phi}{\partial x}(f(x) + g(x)u) + Q(x) + u^T R u. \tag{9}$$

If the plant dynamics are known, one can let the optimal controller be given by:

$$u = -\frac{1}{2} R^{-1} g(x)^T \frac{\partial \phi}{\partial x}^T W_a + e_a(x)$$

where $W_a \in \mathbb{R}^N$ is an unknown parameter. The addition of the parameter W_a ensures that the Hamilton-Jacobi Bellman constraint (9) is linear with respect to the unknown parameter W_c. This is often referred to as an actor-critic approach to reinforcement learning. The solution of this problem requires the estimation of both W_a and W_c subject to an equilibrium constraint $W_a = W_c$. The actor-critic approach yields a very difficult adaptive parameter estimation which requires complex and conservative persistency of excitation conditions. As a result, the tuning and dither signal design is extremely challenging. Furthermore, the corresponding estimation procedures yield local stability properties.

The set-based estimation technique exploits the following parameterization of the Q-function as follows:

$$\begin{aligned}\mathcal{H}_c(W_a, W_c, \phi, x, u) =& (W_a + W_c)^T \frac{\partial \phi}{\partial x}(f(x) + g(x)u) \\ &+ Q(x) + u^T R u + e_c(x).\end{aligned}$$

with nominal state-feedback:

$$u = -\frac{1}{2} R^{-1} g(x)^T \frac{\partial \phi}{\partial x}^T (W_a + W_c) + e_a(x).$$

where $e_c(x)$ and $e_a(x)$ are assumed to be error terms associated with the parameterization of the unknown value function. It is assumed that the error terms are locally Lipschitz on some set $\Omega \subset \mathbb{R}^n$ containing the origin. Throughout this paper, it is assumed that the parameterization is such that $e_c(x) = 0$ and $e_a(x) = 0$. That is, it assumed that the approximation errors are negligible. This assumption can be relaxed but it requires a more complicated analysis that will be addressed in future work.

We first define the approximate error using the Bellman equation:

$$e_H = \mathcal{H}(x^*, u^*) - \left((W_a + \hat{W}_c)^T(L_f\phi + L_g\phi\hat{u}) + Q(x) + \hat{u}^T R\hat{u}\right)$$

where x^* and u^* represents the optimal solution to the optimal control problem where $\mathcal{H}(x^*, u^*) = 0$. The approximate state-feedback is given by:

$$\hat{u} = -\frac{1}{2}R^{-1}g(x)^T\frac{\partial\phi}{\partial x}^T(W_a + \hat{W}_c).$$

Next we pose an estimation approach for $\hat{W}_c$ where it is first assumed that the value of W_a is assumed to be constant. The Bellman error is written as follows:

$$e_H = H(x, u, W_a, W_c) - \left((W_a + \hat{W}_c)^T L_f\phi + Q(x) - \frac{1}{4}(W_a + \hat{W}_c)^T L_g\phi R^{-1}L_g\phi^T(W_a + \hat{W}_c)\right)$$

Expanding, we obtain:

$$\begin{aligned} e_H = H(x, u, W_a, W_c) - \Big((W_a + \hat{W}_c)^T L_f\phi + Q(x) \\ - \frac{1}{4}(W_a)^T L_g\phi R^{-1}L_g\phi^T(W_a)) \\ - \frac{1}{2}(W_a)^T L_g\phi R^{-1}L_g\phi^T(\hat{W}_c)) \\ - \frac{1}{4}(\hat{W}_c)^T L_g\phi R^{-1}L_g\phi^T(\hat{W}_c))\Big) \end{aligned}$$

The first term, can be expanded as:

$$\begin{aligned} H(x, u, W_a, W_c) =& (W_a + W_c)^T L_f\phi + Q(x) \\ &- \frac{1}{4}(W_a)^T L_g\phi R^{-1}L_g\phi^T(W_a) \\ &- \frac{1}{2}(W_a)^T L_g\phi R^{-1}L_g\phi^T(W_c) \\ &- \frac{1}{4}(W_c)^T L_g\phi R^{-1}L_g\phi^T(W_c). \end{aligned}$$

We define the parameter estimation error $\tilde{W}_c = W_c - \hat{W}_c$. The error term e_H is written as:

$$\begin{aligned} e_H =& \tilde{W}_c^T L_f\phi - \frac{1}{2}(W_a)^T L_g\phi R^{-1}L_g\phi^T(\tilde{W}_c) \\ &- \frac{1}{4}(W_c)^T L_g\phi R^{-1}L_g\phi^T(\tilde{W}_c) \\ &- \frac{1}{4}(\tilde{W}_c)^T L_g\phi R^{-1}L_g\phi^T(\hat{W}_c)\Big). \end{aligned}$$

We define

$$\phi_H = L_f\phi - \frac{1}{2}L_g\phi R^{-1}L_g\phi^T(W_a) - \frac{1}{4}L_g\phi R^{-1}L_g\phi^T(\hat{W}_c)$$

and

$$\eta_H = -\frac{1}{4}(W_c)^T L_g\phi R^{-1}L_g\phi^T(\tilde{W}_c).$$

We now rewrite e_H as follows:

$$e_H = \phi^T\tilde{W}_c + \eta_H.$$

We first make the following assumption.

Assumption 1. The true value of the parameter $W_a + W_c$ lies inside a ball of radius z_c centred at W_a denoted by Θ_c.

In this study, we consider the application of a set-based identification technique proposed in Adetola et al. (2014). We first define a variable z such that:

$$\dot{z} = e_H = \tilde{W}_c^T\phi_H + \eta_H$$

Next, we introduce the variables $\hat{z}$ and c, governed by the differential equations:

$$\dot{\hat{z}} = k_W(z - \hat{z}) + c^T\dot{\hat{W}}_c \tag{10}$$

and

$$\dot{c} = -k_W c + \phi_H \tag{11}$$

where k_W is a positive constant to be assigned.

We define the error $e = z - \hat{z}$ with dynamics governed by the differential equations:

$$\dot{e} = \tilde{W}_c^T\phi_H + \eta_H - k_W e - c^T\dot{\hat{W}}_c.$$

Following Adetola et al. (2014), we define the auxiliary variable $\eta = e - c^T\tilde{W}_c$. Their dynamics are given by:

$$\dot{\eta} = -k_W + \eta_H.$$

We then consider the variable $\hat{\eta}$ with dynamics:

$$\dot{\hat{\eta}} = -k_W\hat{\eta} \tag{12}$$

and let $\tilde{\eta} = \eta - \hat{\eta}$ such that:

$$\dot{\tilde{\eta}} = -k_W\tilde{\eta} + \eta_H.$$

We then pose the following parameter update for $\hat{W}_c$:

$$\dot{\Sigma} = cc^T, \tag{13}$$

$$\dot{\hat{W}}_c = \text{Proj}\left(\Sigma^{-1}c(e - \hat{\eta}), \Theta_c\right) \tag{14}$$

with initial conditions $\Sigma(0) = I$, $\hat{W}_c = 0$ where I is the identity matrix and $\text{Proj}(\cdot, \Theta_c)$ denotes the projection algorithm.

We define the function $\mathcal{P} = \hat{W}_c^T\hat{W}_c^T - z_\theta^2$. It is defined as follows:

$$\text{Proj}(\tau, \mathcal{B}_a) = \begin{cases} \tau & \text{if } \mathcal{P} < 0 \text{ or} \\ & \mathcal{P} = 0 \text{ and } \nabla_{\hat{W}_c}\mathcal{P}\tau \leq 0 \\ \left(I - \dfrac{\nabla\mathcal{P}^T\mathcal{P}}{\mathcal{P}\mathcal{P}^T}\right)\tau & \text{otherwise} \end{cases}$$

The algorithm has the following important properties:

(1) It is Lipschitz continuous on Θ_c.
(2) For $\hat{W}_c(0) + W_a \in \Theta_c \Rightarrow \hat{W}_c(t) + W_a \in \Theta_c$, $\forall t \geq 0$.
(3) It fulfills the following inequality:

$$\tilde{W}_c^T\text{Proj}(\tau, \mathcal{B}_a) \geq \tilde{W}_c^T\tau \tag{15}$$

for $W_c \in \Theta_c$ and $\hat{W}_c \in \Theta_c$.

We know pose the Lyapunov function candidate: $V_{\tilde{W}_c} = \tilde{W}_c^T\Sigma\tilde{W}_c$. The time derivative of V along the trajectories of the parameter estimation update is given by:

$$\begin{aligned} \dot{V}_{\tilde{W}_c} =& -2\tilde{W}_c^T\Sigma\dot{\hat{W}}_c + \tilde{W}_c^T cc^T\tilde{W}_c \\ \leq& -\tilde{W}_c^T c(e - \hat{\eta}) + \tilde{W}_c^T cc^T\tilde{W}_c. \end{aligned}$$

We substitute for $\eta = e - c^T\tilde{W}_c$ to obtain:

$$\begin{aligned} \dot{V}_{\tilde{W}_c} \leq& -2(e - \hat{\eta} - \tilde{\eta})^T(e - \hat{\eta}) + (e - \hat{\eta} - \tilde{\eta})^T(e - \hat{\eta} - \tilde{\eta}) \\ \leq& -(e - \hat{\eta})^T(e - \hat{\eta})^2 + \tilde{\eta}^T\tilde{\eta}. \end{aligned}$$

Next we consider the function $V_\eta = \tilde{\eta}^T\tilde{\eta}$. Its derivative is given by:

$$\dot{V}_\eta = -2k_W\tilde{\eta}^T\tilde{\eta} + 2\tilde{\eta}^T\eta_H \leq -k_W\tilde{\eta}^T\tilde{\eta} + \eta_H^T\eta_H.$$

By assumption 1, it follows that $\|W_c\| \leq z_c$ and $\|\tilde{W}_c\| \leq 2z_c$. As a result, we obtain the following inequality:

$$\dot{V}_\eta \leq -k_w\tilde{\eta}^T\tilde{\eta} + \left(\|W_c\|\|L_g\phi L_g\phi^T\|_{\mathcal{F}}\|\tilde{W}_c\|\right)^2$$
$$\leq -k_w V_\eta + 4z_c^4\|L_g\phi L_g\phi^T\|_{\mathcal{F}}^2.$$

where $\|L_g\phi L_g\phi^T\|_{\mathcal{F}}$ denotes the Frobenius of the matrix $L_g\phi L_g\phi^T$.

It follows that if the minimum eigenvalue of Σ is strictly positive and the time varying regressor ϕ_H is bounded, V is a suitable Lyapunov function for the parameter update. This can be stated by the standard persistency of excitation condition.

Assumption 2. The trajectories of the system are such that there exist positive constants, α, β and T such that

$$\alpha I \leq \frac{1}{T}\int_t^{t+T} c(\tau, W_a)c(\tau, W_a)^T d\tau \leq \beta I$$

$\forall t \geq T$ and $\forall W_a \in \Theta_c$.

Next, we invoke a result from Adetola et al. (2014). It is restated with some modifications to address the specific learning task considered in this work.

Lemma 1. Assume that the signals of the system (1) fulfill the persistency of excitation condition as stated in Assumption 2. Then, the parameter estimation scheme (10), (11), (12), (13) and (14) is such that the parameter estimation error converges exponentially to a ball centred at the origin with a radius of $\mathcal{O}(z_{\theta 0}^2)$.

Next we consider a hybrid learning algorithm to identify the constant parameter W_a. This can be done as follows.

We first initialize the uncertainty set with an initial centre $W_a[0]$ and radius $z_c[0]$. The corresponding uncertainty set Θ_c is assumed to contain the true unknown parameter value $W_a[0] + W_c$. As the parameter W_c is updated via (14), we assume that the proposed discrete update yields a pair of sequences $\{W_a[k]\}$ and $\{z_c[k]\}$ associated with a sequence of times $\{t_k\}$ such that

$$W_a(t) = W_a[k],\ t_k \leq t < t_{k+1}. \tag{16}$$

We consider the function V_z as the solution of the differential equation:

$$\dot{V}_z = -(e-\hat{\eta})^T(e-\hat{\eta}) + V_\eta \tag{17}$$
$$\dot{V}_\eta = -k_w V_\eta + z_c[k]^4\|L_g\phi L_g\phi^T\|_{\mathcal{F}}^2 \tag{18}$$

with initial condition, $V_z(0) = 4\lambda_{\max}\left[\Sigma(0)\right](z_c(0))^2$ and $V_\eta(0) = \|\tilde{\eta}(0)\|^2$, respectively. We define the quantity:

$$z_c(t) = \sqrt{\frac{4V_z(t)}{\lambda_{min}[\Sigma(t)]}}. \tag{19}$$

The sequence of times t_k are the times at which

$$z_c(t) \leq z_c[k-1] - \|\hat{W}_c(t)\|. \tag{20}$$

At time t_k, we update the uncertainty set to $W_a[k] = W_a[k-1] + \hat{W}_c(t_k)$ with radius $z_c[k] = z_c(t_k)$. The continuous-time parameter $\hat{W}_c(t)$ is reinitialized to 0. All other quantities are kept at their current value.

This can be summarized by the following algorithm.

Algorithm 1. Beginning from time $t_{i-1} = t_0$, the parameter and set adaptation is implemented iteratively as follows:

1 **Initialize** $z_c(t_{i-1}) = z_c[i-1]$, $\hat{W}_c(t_{i-1}) = 0$, $W_a(t_{i-1}) = W_a[i-1]$, $\hat{\eta}(t_{i-1}) = e(t_{i-1})$, $c(t_{i-1}) = 0$ and $\Theta_c[i-1] = B(W_a[i-1], z_c[i-1]))$.
2 At time t_i, using equations (14) and (19) **perform** the update

$$(W_a[i],\ \Theta[i]) = \begin{cases} \left(W_a[i-1] + \hat{W}_c(t_i),\ \Theta(t_i)\right), \\ \qquad \text{if } z_c(t_i) \leq z_c[i-1] - \|\hat{W}_c(t_i)\| \\ (W_a[i-1],\ \Theta[i-1]), \\ \qquad \text{otherwise} \end{cases} \tag{21}$$

3 **Iterate back** to step 2, **incrementing** $i = i+1$.

As outlined in Adetola et al. (2014), this procedure has three important properties:

(1) $\Theta_c[k+1] \subset \Theta_c[k]\ \forall k$.
(2) $W^* \in \Theta_c[0] \Rightarrow W^* \in \Theta[k]\ \forall k$.

One of the main properties of the proposed approach can be stated as follows.

Theorem 1. Let the trajectories of the system be such that Assumption 2 holds. Let the system be such that Assumptions 1 hold then the set update (19), algorithm 1 and the parameter estimation routine ((10),(11),(12) and (14)) guarantee that the parameter estimates $\hat{\theta}$ converge asymptotically to the true values, θ.

4. PHASOR ESC

In the absence of precise knowledge of the dynamics, the state feedback $\hat{u} = -\frac{1}{2}R^{-1}L_g\phi^T(W_a + W_c^*)$ cannot be implemented explicitly. In existing techniques (see Mehta and Meyn (2009), Bhasin et al. (2013)), the unknown nonlinear system dynamics must be approximated using another affine parameterization.

These approximations can involve very high dimensional neural networks that require considerable exploration and limits the performance of the reinforcement learning algorithms. In this study, we focus on the application of a phasor extremum-seeking control approach to estimate the unknown vector valued function:

$$G(x) = \frac{\partial\phi}{\partial x}g(x) = \left[G_1(x)\ \cdots\ G_N(x)\right]^T$$

We favour the phasor extremum seeking approach since it can provide a direct estimate of $G(x)$. The phasor approximation of $G(x)$ is developed as follows.

One of the characteristics of the phasor approximation is that it requires high frequency signals. In fact, the approximation of $\dot{\phi}$ needs to be interpreted using a time-scale argument. We consider the time-scale transformation $d\tau = \omega dt$ and compute $\dot{\phi}_i$ for the i^{th} element of ϕ in the τ time-scale to obtain:

$$\frac{d\phi_i}{d\tau} = \frac{1}{\omega}(L_f\phi_i + L_g\phi_i\hat{u} + AL_g\phi_i\sin(\tau).$$

If we let $A = \rho\omega$ for some positive constant $\rho > 0$, we therefore obtain:

$$\frac{d\phi_i}{d\tau} = \frac{1}{\omega}(L_f\phi_i + L_g\phi_i\hat{u} + \rho L_g\phi_i\sin(\tau)$$

Using the order notation $\mathcal{O}$, we can write:

$$\frac{d\phi_i}{d\tau} = \mathcal{O}\left(\frac{1}{\omega}\right) + \rho L_g \phi_i \sin(\tau)$$

As a result, we can express $\dot{\phi}_i$ using a phasor approximation of the form:

$$\dot{\phi}_i \approx \beta_{i0} + \alpha_{i1}\sin(\omega t) + \beta_{i1}\cos(\omega t).$$

which is exact up to a term of order $\mathcal{O}\left(\frac{1}{\omega}\right)$.

For each element of the vector $G(x)$, we design a Kalman filter to provide estimations of the time-varying terms β_{i0} α_{i1} and β_{i1}. The phasor estimation dynamics are given by:

$$\dot{\hat{z}}_i = \beta \begin{bmatrix} L_1(\dot{\phi}_i - C\hat{z}_i) + \hat{v}_{1i} \\ L_2(\dot{\phi}_i - C\hat{z}_i) + \hat{v}_{2i} \\ L_3(\dot{\phi}_i - C\hat{z}_i) + \hat{v}_{3i} \end{bmatrix} \tag{22}$$

$$\dot{\hat{v}}_i = \beta\gamma \begin{bmatrix} L_1(\dot{\phi}_i - C\hat{z}_i) \\ L_2(\dot{\phi}_i - C\hat{z}_i) \\ L_3(\dot{\phi}_i - C\hat{z}_i) \end{bmatrix} \tag{23}$$

where $\hat{z}_i = [\hat{z}_{1i},\, \hat{z}_{2i},\, \hat{z}_{3i}]^T$ is the estimate of the vector

$$z = [\beta_{0i},\, \alpha_{i1},\, \beta_{i2}]^T,$$
$$\hat{v}_i = [\hat{v}_{1i},\, \hat{v}_{2i},\, \hat{v}_{3i}]^T$$

and

$$C = [1,\, \sin(\omega t),\, \cos(\omega t)]$$

for $i = 1, \ldots, N$, γ and β are positive constants to be assigned. The observer gain is given by:

$$\begin{aligned} L =& \begin{bmatrix} L_1,\, L_2,\, L_3 \end{bmatrix}^T \\ =& \begin{bmatrix} \beta\ell_1,\ \beta\ell_2\sin(\omega t + \xi),\ \beta\ell_2\cos(\omega t + \xi) \end{bmatrix}^T \end{aligned}$$

with $\xi = 2\tan^{-1}(\phi)$, for some phase angle ϕ to be specified.

5. ROBUST STABILIZATION WITH SET-BASED LEAST SQUARES LEARNING

In this section, we consider the stability of the phasor estimation with the set-based estimation. We first consider the stability of the closed-loop system with state feedback $u = -\frac{1}{2}\hat{G}^T(W_a + \hat{W}_c)$.

We first make the following assumption concerning the existence of a state-feedback.

Assumption 3. The basis function s $\phi(x)$ are chosen such that there exists a W^* and a state-feedback $u = -kL_g\phi^T W^*$ with $k > 0$, a positive constant, and R, a positive definite matrix, that globally asymptotically stabilizes the origin of the closed-loop nonlinear system:

$$\dot{x} = f(x) - kg(x)R^{-1}g(x)^T \frac{\partial \phi}{\partial x}^T W^*.$$

Based on the assumption, one can consider the function $V(x^a) = W^{*T}\phi(x)$ as a candidate Lyapunov for the system. This is clear since the parameter W^* meets the optimality condition:

$$W^{*T}L_f\phi - k(1-k)W^{*T}L_g\phi R^{-1}L_g\phi^T W^* + Q(x) = 0.$$

Thus, for $k = \frac{1}{2}$, one obtains:

$$\begin{aligned} \dot{V} =& W^{*T}L_f\phi - \frac{1}{2}W^{*T}L_g\phi R^{-1}L_g\phi^T W^* \\ =& -Q(x) - \frac{1}{4}W^{*T}L_g\phi R^{-1}L_g\phi^T W^* \leq -Q(x). \end{aligned}$$

Thus, given W^*, the phasor extremum seeking control, the state-feedback, $u = -\frac{1}{2}\hat{G}^T W^*$, yields the following averaged closed-loop:

$$\begin{aligned} \dot{x}^a &= f(x^a) - kg(x^a)\hat{G}^T W^* \\ \frac{d}{dt}\begin{bmatrix} \tilde{z}_i \\ \hat{v}_i \end{bmatrix} &= \beta \begin{bmatrix} -A(\xi) & -I \\ \gamma A(\xi) & 0 \end{bmatrix} \begin{bmatrix} \tilde{z}_i \\ \hat{v}_i \end{bmatrix} \end{aligned} \tag{24}$$

It then follows by the exponential stability of the phasor estimation error dynamics and the optimality of the state-feedback $u = -\frac{1}{2}L_g\phi^T(W_a + W_c^*)$ that the closed-loop system (24) is robustly stable to the parameter estimation error $\tilde{W}_c$ and the phasor estimation error $\tilde{G} = L_g\phi - \hat{G}$.

To avoid the possibility of peaking due to the variations of the estimate $\hat{G}$, one needs to implement a saturated version of estimate of $G = L_g\phi$ to prevent any destabilizing fluctuations. To make this more precise we identify a level set of the function $V = (W_a + W_c^*)^T\phi$, $\Omega_\beta = \{x \in \mathbb{R}^n \,|\, V(x) \leq \beta\}$ and assume that there exists a finite value $M_g = \sup_{x\in\Omega_\beta} \|L_g\phi\|$. As result, the saturation of the estimate such that $\|\hat{G}\| \leq M_g$ is used to define the state-feedback, $u = -\frac{1}{2}\bar{\hat{G}}^T(W_a + W_c^*)$, where $\bar{\hat{G}}$ is a bounded value of the estimate. In this study, we consider the expression:

$$\bar{\hat{G}}_i = M_g \tan^{-1}\left(\frac{\hat{G}_i}{M_g}\right).$$

A Lyapunov stability analysis on the averaged process dynamics can be conducted to confirm the asymptotic stability of the averaged closed-loop system. As a result, one can used a standard average analysis approach to prove the practical asymptotically stability of the closed-loop nonlinear system.

6. SIMULATION STUDY

6.1 Example 1

In this example, we consider the unknown nonlinear system:

$$\begin{aligned} \dot{x}_1 =& -x_1 + x_2, \\ \dot{x}_2 =& -0.5x_1 - 0.5x_2(1 - (\cos(2x_1) + 2)^2) \\ & + (\cos(2x_1) + 2)u \end{aligned}$$

We consider the optimal control problem:

$$J(x_0, u(t)) = \min_{u(t)} \int_0^\infty x(t)^T Q^T x(t) + u(t)^T R u(t) dt$$

where $Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $R = 1$. The basis functions are chosen as: $\phi(x) = [x_2^2,\, x_1x_2,\, x_1^2]$.

The vector of derivatives $\dot{\phi}$ is estimated using a high-pass filter with filter parameter $\omega_h = 1000$. We consider a dither

frequency $\omega = 200$ with amplitude $A = 15$. The phasor estimator is tuned such that $l_1 = 1000$ and $l_2 = 100$ and $\xi = \pi/6$. The tuning constants of the parameter estimation routine are given by $\gamma_1 = 1$, $\gamma_2 = 0.2$. The initial conditions are $x(0) = [10, 10]^T$, $\hat{W}_c(0) = \hat{W}_a(0) = [10.251]^T$, with all other variables starting at zero. No external dither signal is required

$$u = -k\hat{W}_a^T \hat{G}(x) + d(t)$$

where $d(t) = 5\sin(2t)$.

The value function for this optimal control problem is:

$$V^*(x) = x^T \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 1 \end{bmatrix} x$$

which can be optimally represented over the functions in the vector $\phi(x)$.

The simulation results are shown in Figure 1-2. Figure 1 shows the state variables (x_1 , x_2) and the input (u) trajectories resulting from the learning algorithm. The parameter estimates $\hat{W}_c$ (full lines) are shown in Figure 2 along with the actual value W_c (dashed line) corresponding to the true value function $V^*(x)$. The proposed technique approximates the value function of this nonlinear optimal control problem effectively.

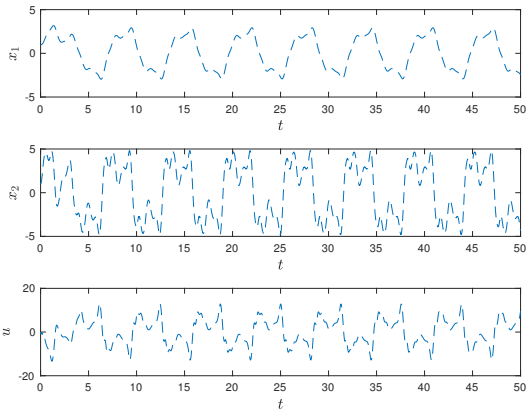

Fig. 1. The state variables (x_1 , x_2) and input (u) trajectories for Example 6.1.

7. CONCLUSION

This study presents a new Q-learning technique for the approximation of output controller for a class of unknown nonlinear systems. The approach incorporates two new techniques for the approximation of the value function of an infinite horizon optimal control problems. A set-based estimation is proposed for the estimation of the value function. A phasor estimation approach is used to circumvent the lack of knowledge of the nonlinear dynamics. The technique is shown to provide an effective approximation of the value function and the optimal controller that solves a user-defined nonlinear optimal control problem.

REFERENCES

Adetola, V., Guay, M., and Lehrer, D. (2014). Adaptive estimation for a class of nonlinearly parameterized dynamical systems. *IEEE Transactions on Automatic Control*, 59(10), 2818–2824.

Atta, K.T., Johansson, A., and Gustafsson, T. (2015). Extremum seeking control based on phasor estimation. *Systems & Control Letters*, 85, 37–45.

Barto, A.G., Sutton, R.S., and Anderson, C.W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5), 834–846. doi: 10.1109/TSMC.1983.6313077.

Bertsekas, D.P. and Tsitsiklis, J.N. (1995). Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, 560–564. IEEE.

Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K.G., Lewis, F.L., and Dixon, W.E. (2013). A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica*, 49(1), 82–92.

Bradtke, S.J., Ydstie, B.E., and Barto, A.G. (1994). Adaptive linear quadratic control using policy iteration. In *American Control Conference, 1994*, volume 3, 3475–3479 vol.3. doi:10.1109/ACC.1994.735224.

Busoniu, L., Babuska, R., De Schutter, B., and Ernst, D. (2010). *Reinforcement learning and dynamic programming using function approximators*, volume 39. CRC press.

Mehta, P. and Meyn, S. (2009). Q-learning and pontryagin's minimum principle. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, 3598–3605. IEEE.

Sutton, R.S., Barto, A.G., and Williams, R.J. (1992). Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems*, 12(2), 19–22. doi: 10.1109/37.126844.

Sutton, R.S. and Barto, A.G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

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

Watkins, C.J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4), 279–292.

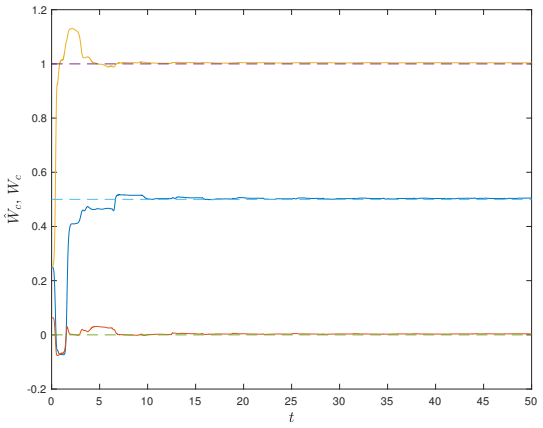

Fig. 2. Plot of the parameter estimates $\hat{W}_c$ (full linea) and the actual value W_c (dashed lines) for Example 6.1.