Difan Tang, Lei Chen, Zhao Feng Tian and Eric Hu
Modified value-function-approximation for synchronous policy iteration with single-critic configuration for nonlinear optimal control
International Journal of Control, 2019; OnlinePubl:1-13

Modified value-function-approximation for synchronous policy iteration with single-critic configuration for nonlinear optimal control

ARTICLE HISTORY
Compiled April 23, 2019

ABSTRACT
This study proposes a modified value-function-approximation (MVFA) and investigates its use under a single-critic configuration based on neural networks (NNs) for synchronous policy iteration (SPI) to deliver compact implementation of optimal control online synthesis for control-affine continuous-time nonlinear systems. Existing single-critic algorithms require stabilising critic tuning laws while eliminating actor tuning. This paper thus studies alternative single-critic realisation aiming to relax the needs for stabilising mechanisms in the critic tuning law. Optimal control laws are determined from the Hamilton-Jacobi-Bellman equality by solving for the associated value function via SPI in a single-critic configuration. Different from other existing single-critic methods, an MVFA is proposed to deal with closed-loop stability during online learning. Gradient-descent tuning is employed to adjust the critic NN parameters in the interests of not complicating the problem. Parameters convergence and closed-loop stability are examined. The proposed MVFA-based approach yields an alternative single-critic SPI method with uniformly ultimately bounded closed-loop stability during online learning without the need for stabilising mechanisms in the critic tuning law. The proposed approach is verified via simulations.

KEYWORDS
Adaptive dynamic programming; approximate dynamic programming; neural networks; nonlinear control; optimal control; policy iteration

1. Introduction

Nonlinear optimal control generally involves the determination of control laws that minimise the associated performance cost, where the Hamilton-Jacobi-Bellman (HJB) equality (Bellman, 1957) or its nonlinear variations are to be solved, or where an inverse approach without solving the HJB equation (Lopez, Sanchez, Alanis, & Rios, 2017) may apply. In our study, the discussion is focused on the former, where the HJB equality and its variants, being partial differential equations that are nonlinear, are difficult to be solved analytically. Practical methods to solve the HJB equation and its variants are provided through approximation methods, one class of which is the widely studied adaptive/approximate dynamic programming (ADP) (Werbos, 1974). ADP techniques are basically iterative approaches built upon the concept of reinforcement learning (Sutton & Barto, 1998), which approximates optimal control laws as well as corresponding value functions through policy evaluation and improvement, where a 'policy' is referred to as a control law. Some good reviews are provided by F.-Y. Wang, Zhang, and Liu (2009), Z.-P. Jiang and Jiang (2013), and D. Wang, He, and Liu (2017a). To implement the ADP, the value function in the HJB equation needs to be properly structured, and neural networks (NNs) are ideal candidates given their universal approximation properties (Hornik, Stinchcombe, & White, 1989).

Offline ADP has been an effective and useful tool for handling optimal control in various challenging problems, including nonaffine systems (Luo, Liu, Huang, & Wang, 2016; Mu, Wang, & He, 2017; D. Wang, Liu, Wei, Zhao, & Jin, 2012), actuator saturation (Abu-Khalaf & Lewis, 2005; Heydari & Balakrishnan, 2013; Luo, Wu, Huang, & Liu, 2015), unknown system dynamics (Li, Modares, Chai, Lewis, & Xie, 2017; Luo et al., 2016, 2015; Mu et al., 2017; Mu, Wang, & He, 2018; D. Wang & Liu, 2013; D. Wang et al., 2012; Wei, Lewis, Sun, Yan, & Song, 2017; Zhao, Xia, & Wang, 2015), fixed final time (Heydari & Balakrishnan, 2013), finite approximation error (Wei, Wang, Liu, & Yang, 2014), finite horizon (Mu et al., 2018), algorithm simplification (Heydari, 2014; Heydari & Balakrishnan, 2013; D. Wang & Liu, 2013), optimal tracking (Luo et al., 2016), non-zero initial condition for value iteration (Wei, Liu, & Lin, 2016), and extension to multi-agent system applications (Li et al., 2017).

With increasing demands on synthesising optimal controllers in real time, online ADP has been receiving intensive research attention. Online ADP, in contrast to offline methods, features real-time synthesis of optimal control policies for dynamic systems. The iteration procedures performed on a regular- or irregular-time-interval basis, where the cost function corresponding to an admissible control being approximated undergoes evaluation before the next iteration commences, can be characterised as being *sequential*. These algorithms collect real-time data prior to batch processing for policy evaluation and policy update at each discrete iteration under either continuous-time setting (Feng, Zhang, Luo, & Zhang, 2015; Y. Jiang & Jiang, 2014, 2015; Liu, Yang, & Li, 2013; Vrabie & Lewis, 2009) or in discrete-time domain (Al-Tamimi, Lewis, & Abu-Khalaf, 2008; Feng et al., 2015; Kiumarsi, Lewis, & Levine, 2015; Škach, Kiumarsi, Lewis, & Straka, 2018; Wei & Liu, 2014). The study by Vamvoudakis and Lewis (2010) proposes an attractive ADP algorithm, termed as *synchronous policy iteration* (SPI), where policy evaluation and policy update are implemented continuously in time and simultaneously. The SPI theory framework initiated by Vamvoudakis and Lewis (2010) has been enormously enriched by latest advances in dealing with faster convergence (Bhasin et al., 2013), actuator saturation (Huang, Wang, & Liu, 2017; Kiumarsi & Lewis, 2015; Modares & Lewis, 2014; Modares, Lewis, & Naghibi-Sistani, 2013, 2014; Modares, Naghibi Sistani, & Lewis, 2013; Yang, Liu, & Wang, 2014), completely unknown dynamics with unknown nonlinear structures (Liu, Huang, Wang, & Wei, 2013; Yang et al., 2014), unknown affine nonlinear systems (Lv, Na, & Ren, 2017; Lv, Na, Yang, Wu, & Guo, 2016; Modares, Lewis, & Naghibi-Sistani, 2013; Na & Herrmann, 2014; Song, Lewis, Wei, & Zhang, 2016; D. Wang, Liu, Zhang, & Zhao, 2016; Zhong, He, Wang, & Ni, 2018), partially unknown dynamics (Bhasin et al., 2013; Kiumarsi & Lewis, 2015; Modares & Lewis, 2014; Modares et al., 2014; Vamvoudakis, Vrabie, & Lewis, 2014), multi-agent systems (Heydari & Balakrishnan, 2014; H. Jiang & He, 2018; Luy, 2018), optimal tracking (Kiumarsi & Lewis, 2015; Modares & Lewis, 2014; Na & Herrmann, 2014), relaxation of persistent-excitation condition (Modares et al., 2014), exponential convergence driven directly by estimation error assuming known ideal parameters rather than being driven by the HJB error (Lv et al., 2017, 2016; Na & Herrmann, 2014), algorithm simplification (Huang et al., 2017; Liu, Huang, et al., 2013; Liu, Wang, Wang, Li, & Yang, 2014; Luy, 2018; Lv et al., 2017, 2016; Na & Herrmann, 2014; D. Wang, He, & Liu, 2017b; D. Wang, Liu, Li, & Ma, 2014; D. Wang, Mu, Yang, & Liu, 2017; Zhang, Cui, & Luo, 2013), and disturbances and uncertainties (Huang et al., 2017; Liu et al., 2014; Lv et al., 2017; Song et al., 2016; Vamvoudakis & Lewis, 2012; D. Wang et al., 2014, 2016).

For stabilisation purpose, most SPI schemes implement separate NNs for the critic and actor, respectively, each dynamically tuned with a different learning law. Specif- 1.5

ically, actor tuning laws generally contain stabilising terms derived from Lyapunov stability analysis. To simplify SPI implementation and reduce computational load, there have been efforts on single-critic approaches where the same NN is used for both components with the critic NN weights directly passed on to the actor NN (Huang et al., 2017; Liu, Huang, et al., 2013; Liu et al., 2014; Luy, 2018; Lv et al., 2017, 2016; Na & Herrmann, 2014; D. Wang, He, & Liu, 2017b; D. Wang et al., 2014; D. Wang, Mu, et al., 2017; Zhang et al., 2013). Further improvements are seen in event-based methods based on the single-critic configuration (D. Wang, He, & Liu, 2017b; D. Wang, Mu, et al., 2017), where the data needed for online learning are reduced. The instability resulted from direct simplification of the actor-critic configuration is recognised in Liu, Huang, et al. (2013), and critic-NN initial weights need to be determined carefully by trial-and-error. Guaranteed stability can be achieved by introducing a stabilising mechanism to the critic tuning law (Huang et al., 2017; Liu et al., 2014; Luy, 2018; Lv et al., 2017, 2016; Na & Herrmann, 2014; D. Wang, He, & Liu, 2017b; D. Wang et al., 2014; Zhang et al., 2013). The stabilising mechanism is generally a stabilising term derived on the basis of Lyapunov stability, either conditionally activated upon instability being detected (Huang et al., 2017; Liu et al., 2014; Luy, 2018; D. Wang et al., 2014; Zhang et al., 2013), or continuously in effect throughout online learning (Lv et al., 2017, 2016; Na & Herrmann, 2014; D. Wang, He, & Liu, 2017b). It is interesting to note that the SPI schemes in the aforementioned studies share a common form of value function approximation (VFA) with an NN of standard structure directly employed. The question is: Can a different form of VFA deliver alternative realisation of the single-critic configuration for SPI without introducing additional stabilising mechanisms in the NN tuning law?

Therefore, as our major contributions, this study proposes a modified value-function-approximation (MVFA) and study its feasibility and efficacy as an alternative approach under the single-critic configuration. Specifically, closed-loop stability are investigated.

In the remainder of the paper: Section 2 introduces the problem under discussion together with some preliminaries; Section 3 proposes an MVFA for alternative realisation of the single-critic configuration for SPI; Section 4 analyses overall closed-loop stability during online learning; Section 5 gives two simulation examples. Section 6 draws conclusions.

2. Problem and preliminaries

2.1. Problem

The following control-affine nonlinear systems in continuous-time domain is considered:

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

where $\boldsymbol{x} \in \mathbb{R}^{n_x}$ contains system states of dimension n_x, $\boldsymbol{x}(0) = \boldsymbol{x}_0$, with $\boldsymbol{x}_0$ being a vector containing the initial states; $\boldsymbol{u} \in \mathbb{R}^{n_u}$ collects control inputs of dimension n_u; $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^{n_x}$ refers to internal dynamics of the system; $\boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}^{n_x \times n_u}$ denotes distribution dynamics of control inputs.

Assumption 1. *For the system as in (1), there is* $\boldsymbol{f}(0) = 0$. *Given a set* $\boldsymbol{\Omega} \subseteq \mathbb{R}^{n_x}$ *including zero, equation (1) is Lipschitz continuous with respect to* $\boldsymbol{\Omega}$, *and there exist*

admissible control $\boldsymbol{u} \in \boldsymbol{\Xi}(\boldsymbol{\Omega})$ *that can stabilise (1).* $\boldsymbol{f}(\boldsymbol{x})$ *as well as* $\boldsymbol{g}(\boldsymbol{x})$ *are assumed known.*

Assumption 2. *There exist* $\|\boldsymbol{f}(\boldsymbol{x})\| \leq b_f \|\boldsymbol{x}\|$ *with constant* $b_f \in \mathbb{R}^+$ *and* $\|\boldsymbol{g}(\boldsymbol{x})\| \leq b_g$ *with constant* $b_g \in \mathbb{R}^+$ *(Modares et al., 2014; Modares, Naghibi Sistani, & Lewis, 2013; Vamvoudakis & Lewis, 2010).*

A proper control law $\boldsymbol{u}$ is desired to minimise

$$V(\boldsymbol{x}_0) = \int_0^\infty \left[\bar{Q}(\boldsymbol{x}(t)) + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{u}\right] dt, \tag{2}$$

which is also known as a cost function with a positive-definite function $\bar{Q}(\boldsymbol{x})$ and symmetric positive-definite weighting $\boldsymbol{R} \in \mathbb{R}^{n_u \times n_u}$.

Definition 1 (Admissible control)**.** Given continuously differentiable control $\boldsymbol{u}(\boldsymbol{x}) \in \boldsymbol{\Psi}(\boldsymbol{\Omega})$ with initial condition $\boldsymbol{u}(0) = 0$, if on $\boldsymbol{\Omega}$ it stabilises system (1) and if the cost $V(\boldsymbol{x}_0)$, $\forall \boldsymbol{x}_0 \in \boldsymbol{\Omega}$, as given in (2) is finite, then the control is considered as being admissible (Beard, Saridis, & Wen, 1997).

2.2. Continous-time HJB equation

If $V \in C^1$, differentiating (2) yields

$$\bar{Q}(\boldsymbol{x}) + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{u} + (\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u})^{\mathrm{T}} \nabla V = 0, \tag{3}$$

with $V(0) = 0$ and $\nabla V \triangleq \frac{\partial V(\boldsymbol{x})}{\partial \boldsymbol{x}} \in \mathbb{R}^{n_x}$.

The control that minimises (2) for the same initial conditions is deemed optimal and denoted as $\boldsymbol{u}^*$. The associated cost is $V^* = \min(V)$ for $\boldsymbol{u} \in \boldsymbol{\Xi}(\boldsymbol{\Omega})$ and generally known as the 'value function'. Specifically,

$$\boldsymbol{u}^* = -\frac{1}{2} \boldsymbol{R}^{-1} \boldsymbol{g}^{\mathrm{T}} \nabla V^*, \tag{4}$$

with which there is

$$\bar{Q} + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{u}^* + (\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^*)^{\mathrm{T}} \nabla V^* = 0, \tag{5}$$

with $V^*(0) = 0$, which then gives the following Hamilton-Jacobi-Bellman (HJB) equation:

$$-\frac{1}{4} \nabla V^{*\mathrm{T}} \boldsymbol{g} \boldsymbol{R}^{-1} \boldsymbol{g}^{\mathrm{T}} \nabla V^* + \nabla V^{*\mathrm{T}} \boldsymbol{f} + \bar{Q} = 0, \tag{6}$$

with $V^*(0) = 0$.

Remark 1. Note that $\boldsymbol{u}$ in (3) can be any admissible control, and there exists a corresponding cost V as in (2) that makes (3) hold. However, (5) is a special case of (3) where $\boldsymbol{u}$ is associated with V through (4). A residual error arises to the right of (5) and (6) if the condition of $V^* = \min(V)$ for $\boldsymbol{u} \in \boldsymbol{\Xi}(\boldsymbol{\Omega})$ is unsatisfied.

2.3. Policy iteration

To analytically determine $V^*(\boldsymbol{x})$ from the nonlinear HJB equation has been known difficult. Instead, $V^*(\boldsymbol{x})$ can be obtained through an iterative procedure termed as 'policy iteration' (Sutton & Barto, 1998), which requires $V^*(\boldsymbol{x})$ being appropriately structured and successively approximated (Saridis & Lee, 1979), basically involving two steps in a 'actor-critic' configuration:

- The 'critic' for policy evaluation: using (3) to evaluate $V_{(i)}$ resulted from $\boldsymbol{u}_{(i)}$. This is to solve for $V_{(i)}$ from

$$\bar{Q}(\boldsymbol{x}) + \boldsymbol{u}_{(i)}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{u}_{(i)} + (\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}_{(i)})^{\mathrm{T}} \nabla V_{(i)} = 0, \tag{7}$$

  with $V_{(i)}(0) = 0$.
- The 'actor' for policy improvement: implementing updated control, which is

$$\boldsymbol{u}_{(i+1)} = -\frac{1}{2} \boldsymbol{R}^{-1} \boldsymbol{g}^{\mathrm{T}} \nabla V_{(i)}. \tag{8}$$

The iteration procedure begins with $\boldsymbol{u}_{(0)}$ which is an initial admissible control, and proceeds with the above two iterative steps until reaching convergence at V^* and $\boldsymbol{u}^*$ or proximity to V^* and $\boldsymbol{u}^*$. It is worth emphasising that for synchronous policy iteration (SPI), the procedure performs continuously in time, and the above two steps take place simultaneously (Vamvoudakis & Lewis, 2010). The subscript '(i)' in $V_{(i)}$ and $\boldsymbol{u}_{(i)}$ are unnecessary in the SPI case. However, for ease of explanation of SPI at an infinitesimal time step, these subscripts are used, only to indicate a general time step being considered rather than iteration number.

Remark 2. In terms of the single-critic configuration, an actor component is still necessary for a complete policy iteration procedure including SPI. The term 'single-critic' refers to the case where the separate tuning for the actor component is eliminated in comparison to the general actor-critic structure in which both of the actor and critic components require individual tuning.

3. Modified single-critic configuration

3.1. Modified value function approximation

Analytically obtaining $V_{(i)}(\boldsymbol{x})$ from (7) is difficult, and hence implementing policy iteration requires proper approximation of the solution. Neural networks (NNs), with universal approximation properties (Hornik et al., 1989), can be used for this purpose. Different from other existing studies that use a common form of NN-based representation for approximating the value function, in this paper a modified value-function-approximation (MVFA) is proposed, being:

$$V^* = \frac{1}{2} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{x} + \boldsymbol{W}^{*\mathrm{T}} \boldsymbol{\Phi} + \varepsilon, \tag{9}$$

where hidden-layer neurons are contained in $\boldsymbol{\Phi} \in \mathbb{R}^{n_n}$, with ideal NN weights being $\boldsymbol{W}^* \in \mathbb{R}^{n_n}$; $\boldsymbol{P} \in \mathbb{R}^{n_x \times n_x}$ is an additional parameter matrix that is diagonal and positive-definite; the error of approximation is denoted by $\varepsilon \in \mathbb{R}$.

Accordingly, there is

$$\nabla V^* = \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^* + \boldsymbol{P}\boldsymbol{x} + \nabla\varepsilon, \tag{10}$$

with $\bar{\nabla}\boldsymbol{\Phi} = \nabla\boldsymbol{\Phi}^{\mathrm{T}} = \left[\frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{x}}\right]^{\mathrm{T}} \in \mathbb{R}^{n_n \times n_x}$ and $\nabla\varepsilon = \frac{\partial\varepsilon}{\partial\boldsymbol{x}} \in \mathbb{R}^{n_x}$.

Remark 3. The discussion in Section 1 has revealed that VFA in existing methods takes a common NN-based representation, the convergence of which in online learning necessitates separate actor tuning or stabilising mechanisms in critic tuning laws for stabilisation. Differently in this study, the proposed MVFA features an auxiliary term in addition to the standard structure of an NN. The advantages of introducing the auxiliary term is to be discussed in the remainder of this paper.

Remark 4. The hidden-layer neurons in $\boldsymbol{\Phi}$ are nonlinear activation functions, which can be obtained by applying Weierstrass approximation using high-order polynomials (Finlayson, 1972). The resulting activation functions are the individual terms of a polynomial of specified order with the NN inputs as variables.

Assumption 3. *There exist inequalities* $\left\|\bar{\nabla}\boldsymbol{\Phi}\right\| \leq b_\phi \left\|\boldsymbol{x}\right\|$ *for* $b_\phi \in \mathbb{R}^+$ *and* $\left\|\nabla\varepsilon\right\| \leq b_\varepsilon \left\|\boldsymbol{x}\right\|$ *for* $b_\varepsilon \in \mathbb{R}^+$*, where* b_ϕ *and* b_ε *are constants.*

3.2. Single-critic structure and tuning

On considering the 'Policy Evaluation' step only (i.e., a control law remains fixed for evaluation), the associated cost function $V_{(i)}$ takes

$$V_{(i)} = \boldsymbol{W}_{(i)}^{\mathrm{T}}\boldsymbol{\Phi} + \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x} + \varepsilon_{(i)}, \tag{11}$$

with its gradient being

$$\nabla V_{(i)} = \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}_{(i)} + \boldsymbol{P}\boldsymbol{x} + \nabla\varepsilon_{(i)}, \tag{12}$$

with $\boldsymbol{W}_{(i)}$ being NN ideal weights that approximate $V_{(i)}$ with the least error $\varepsilon_{(i)}$.

Remark 5. Technically speaking, $V_{(i)}$ in (11) and V in (2) are equal only in terms of value, given the same initial conditions and the same control policy, but different in structure. V in (2) is structured to give physical interpretation of cost while $V_{(i)}$ in (11) is specially constructed for mathematical approximation. The term $\boldsymbol{W}_{(i)}^{\mathrm{T}}\boldsymbol{\Phi}$ in (11) is not equal to $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}$ in (2) but includes the information of $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}$, since the set $\boldsymbol{\Phi}$ contains activation functions in polynomial forms consisting of both $\boldsymbol{x}$ and $\boldsymbol{u}$.

Remark 6. The discussion at this stage only considers the case of approximating the cost function for a known control policy $\boldsymbol{u}_{(i)}$. That is, $\boldsymbol{u}_{(i)}$ is known and not approximated by NN. The NN used at this stage only approximates the cost function associated with the known control $\boldsymbol{u}_{(i)}$.

Using an estimate $\hat{\boldsymbol{W}}_{(i)}$ to replace $\boldsymbol{W}_{(i)}$ in (11) and (12) gives

$$\hat{V}_{(i)} = \hat{\boldsymbol{W}}_{(i)}^{\mathrm{T}}\boldsymbol{\Phi} + \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x}, \tag{13}$$

$$\nabla \hat{V}_{(i)} = \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{W}}_{(i)} + \boldsymbol{P}\boldsymbol{x}, \tag{14}$$

and

$$\bar{Q} + \boldsymbol{u}_{(i)}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u}_{(i)} + (\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}_{(i)})^{\mathrm{T}}\nabla \hat{V}_{(i)} = e_1, \tag{15}$$

where $\hat{V}_{(i)}(0) = 0$, and e_1 is the error arises as a result (as commented in Remark 1 and to be discussed in Section 4).

To minimise e_1 so that $\hat{\boldsymbol{W}}_{(i)} \to \boldsymbol{W}_{(i)}$, gradient-descent tuning is adopted, by considering the quadratic error function

$$E = \frac{1}{2}e_1^2. \tag{16}$$

This yields

$$\dot{\hat{\boldsymbol{W}}}_{(i)} = -\kappa_1 \boldsymbol{D}_1 \frac{\partial E}{\partial \hat{\boldsymbol{W}}_{(i)}} = -\frac{a}{\sqrt{{\boldsymbol{\sigma}_{(i)}}^{\mathrm{T}}\boldsymbol{\sigma}_{(i)} + 1}}\boldsymbol{D}_1\boldsymbol{\sigma}_{(i)}e_1, \tag{17}$$

where $\kappa_1 = \frac{a}{\sqrt{\boldsymbol{\sigma}_{(i)}^{\mathrm{T}}\boldsymbol{\sigma}_{(i)}+1}}$ is added for normalisation, with $a \in \mathbb{R}^+$ being a scalar learning rate and $\boldsymbol{\sigma}_{(i)} = \bar{\nabla}\boldsymbol{\Phi}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}_{(i)})$; $\boldsymbol{D}_1$ is an auxiliary term added to adjust contribution of individual state to tuning, and $\boldsymbol{D}_1 = \mathrm{diag}(\boldsymbol{D}_2\boldsymbol{D}_3)$, with $\boldsymbol{D_2} \in \mathbb{R}^{n_n \times n_x}$ being a constant matrix related to $\bar{\nabla}\boldsymbol{\Phi}(\boldsymbol{x})$ with its element $\boldsymbol{D}_{2(jk)} \in \mathbb{B}$, $(j = 1, 2, \cdots, n_n; k = 1, 2, \cdots, n_x)$, and $\boldsymbol{D}_3 \in \mathbb{R}^{n_x \times 1}$ being a weighting vector.

Specifically, the constant matrix $\boldsymbol{D}_2$, in connection with the expression of every single element of $\bar{\nabla}\boldsymbol{\Phi}(\boldsymbol{x})$, namely, $\bar{\nabla}\boldsymbol{\Phi}_{(jk)}(\boldsymbol{x})$, is given in the following form:

$$\boldsymbol{D}_{2(jk)} = \begin{cases} 0 & \text{if } \bar{\nabla}\boldsymbol{\Phi}_{(jk)}(\boldsymbol{x}) = 0,\ \forall \boldsymbol{x} \neq 0, \\ 1 & \text{if } \bar{\nabla}\boldsymbol{\Phi}_{(jk)}(\boldsymbol{x}) \neq 0,\ \forall \boldsymbol{x} \neq 0. \end{cases}$$

Similarly, for the complete synchronous policy iteration (SPI), the ideal weights $\boldsymbol{W}^*$ are unknown and should be determined so that (9) approximates a target value function. With $\hat{\boldsymbol{W}}$ being the estimated weights, the approximated value function and its gradient become

$$\hat{V} = \hat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\Phi} + \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x} \tag{18}$$

and

$$\nabla \hat{V} = \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{W}} + \boldsymbol{P}\boldsymbol{x}, \tag{19}$$

respectively, and the associated control is given by

$$\hat{\boldsymbol{u}} = -\frac{1}{2}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla \hat{V}. \tag{20}$$

Note the absence of the subscript '(i)' in (18) and (19) for complete SPI, which are different from (13) and (14) corresponding to a fixed control law at a general

infinitesimal time step for 'Policy Evaluation' only.

In the SPI case involving the single-critic structure with (18) and direct implementation of (20), there is

$$\bar{Q} + \hat{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{R}\hat{\boldsymbol{u}} + (\boldsymbol{f} + \boldsymbol{g}\hat{\boldsymbol{u}})^{\mathrm{T}}\nabla\hat{V} = e_2, \tag{21}$$

where $\hat{V}(0) = 0$, and e_2 is the resulting approximation error as commented in Remark 1 (details to be given in Section 4).

To minimise e_2 so that $\hat{\boldsymbol{W}} \to \boldsymbol{W}^*$, equation (17) is modified as

$$\dot{\hat{\boldsymbol{W}}} = -\frac{a}{\sqrt{\boldsymbol{\sigma}^{\mathrm{T}}\boldsymbol{\sigma} + 1}}\boldsymbol{D}_1\boldsymbol{\sigma}e_2 = -\kappa_2\boldsymbol{D}_1\boldsymbol{\sigma}e_2, \tag{22}$$

where $\kappa_2 = \frac{a}{\sqrt{\boldsymbol{\sigma}^{\mathrm{T}}\boldsymbol{\sigma}+1}}$, and $\boldsymbol{\sigma} = \bar{\nabla}\boldsymbol{\Phi}(\boldsymbol{f} + \boldsymbol{g}\hat{\boldsymbol{u}})$, with $\boldsymbol{D}_1$, $\boldsymbol{D_2}$, and $\boldsymbol{D}_3$ defined the same as in (17).

It now gives a single-critic structure consisting of critic tuning only, without additional stabilising mechanisms in the tuning law (22).

Remark 7. For conventional VFA as in the SPI pioneer work of Vamvoudakis and Lewis (2010) (also commonly used in other studies discussed in Section 1), it has been known that closing the loop by directly passing $\hat{\boldsymbol{W}}$ on to the actor NN can lead to instability issues during online learning without any stabilising mechanism. This is because in these cases some intermediate values along the evolution path of $\hat{\boldsymbol{W}}$ may not necessarily yield admissible intermediate control policies that satisfy $\frac{d\hat{V}}{dt} < 0$.

Remark 8. Compared with the existing single-critic approaches with conventional VFA and stabilising critic tuning laws, the proposed method with MVFA also differs in that the critic tuning law does not need to be stabilising, allowing the use of simpler tuning laws. Accordingly, in this paper the critic tuning based on traditional yet simple gradient descent is used without additional stabilising mechanisms in the tuning law. Closed-loop stability is to be investigated next in Section 4.

4. Convergence and stability analysis

4.1. Policy evaluation

Similar to most adaptive control problems that require online tuning of parameters (Ioannou & Sun, 1996), proper convergence of NN parameters in this paper also relies on the persistence of excitation (PE) condition to ensure sufficiently rich training set being obtained.

Definition 2 (Persistence of Excitation)**.** A bounded vector signal $\boldsymbol{z}(t)$ is considered to be persistently excited (PE) if

$$\mu_1\boldsymbol{I} \preceq \int_{t_0}^{t_0+t_d} \boldsymbol{z}(\tau)\boldsymbol{z}(\tau)^{\mathrm{T}}d\tau \preceq \mu_2\boldsymbol{I}; \quad \forall t_0 \geq 0,$$

where $\mu_1 \in \mathbb{R}^+$, $\mu_2 \in \mathbb{R}^+$, $t_d \in \mathbb{R}^+$, and $\boldsymbol{I}$ is an identity matrix (Ioannou & Sun, 1996).

In addition, the stability of a linear time-varying system as given by the lemma

below, is to be used in the stability analysis that follows.

Lemma 1. *For a given system being linear and time-varying in the form of*

$$\dot{\boldsymbol{x}} = -\boldsymbol{h}(t)\boldsymbol{h}^{\mathrm{T}}(t)\boldsymbol{x}, \tag{23}$$

where vector $\boldsymbol{x}$ contains system states, its origin is exponentially stable if vector $\boldsymbol{h}(t)$ satisfies the condition of PE (Ioannou & Sun, 1996).

Assumption 4. *During online tuning, states $\boldsymbol{x}(t)$ of the system (1) satisfy the PE condition.*

The following theorem presents the convergence property of *Policy Evaluation* with the MVFA under the tuning given by (17).

Theorem 1. *Let (11) approximate the cost function (2) corresponding to a given admissible control $\boldsymbol{u}_{(i)}$. Under Assumptions 1, 3, and 4, and the tuning algorithm (17), the error $\tilde{\boldsymbol{W}}_{(i)} = \boldsymbol{W}_{(i)} - \hat{\boldsymbol{W}}_{(i)}$ from NN weights estimation converges to a residual set $\boldsymbol{c}_{\tilde{W}}$ exponentially, and $\|\boldsymbol{c}_{\tilde{W}}\| \leq b_{\tilde{W}}$ for a finite scalar $b_{\tilde{W}} \in \mathbb{R}^+$ with $b_{\tilde{W}} \to 0$ as $n_n \to \infty$.*

Proof. Comparing (7) and (15), with (12) substituted for $\nabla V_{(i)}$, and with (14) substituted for $\nabla \hat{V}_{(i)}$, yields

$$e_1 = -\tilde{\boldsymbol{W}}_{(i)}^{\mathrm{T}}\boldsymbol{\sigma}_{(i)} + \epsilon_1, \tag{24}$$

where $\boldsymbol{\sigma}_{(i)} = \bar{\nabla}\boldsymbol{\Phi}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}_{(i)})$, and $\epsilon_1 = -\nabla\varepsilon_{(i)}^{\mathrm{T}}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}_{(i)})$.

As can be seen from (24), if $\varepsilon_{(i)}(\boldsymbol{x}) = 0$ for any $\boldsymbol{x} \neq 0$, then $\epsilon_1 = 0$. For the case of $\varepsilon_{(i)}(\boldsymbol{x}) \neq 0$, it is easy to see that $\epsilon_1 \leq b_{\epsilon_1}$ for $b_{\epsilon_1} \in \mathbb{R}^+$, given Assumption 3 and $(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}_{(i)})$ as well as $\boldsymbol{x}$ being bounded under Assumption 1. Since $\varepsilon_{(i)} \to 0$ and $\nabla\varepsilon_{(i)} \to 0$ given proper activation functions with sufficiently large n_n (Finlayson, 1972), it is straightforward to show that $\epsilon_1 \to 0$ when $n_n \to \infty$.

By using (17) and (24), we have the time derivative of $\tilde{\boldsymbol{W}}_{(i)}$

$$\dot{\tilde{\boldsymbol{W}}}_{(i)} = -a\boldsymbol{D}_1\boldsymbol{\sigma}_{na(i)}\boldsymbol{\sigma}_{na(i)}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{(i)} + a\boldsymbol{D}_1\boldsymbol{\sigma}_{nb(i)}\epsilon_1, \tag{25}$$

where $\boldsymbol{\sigma}_{na(i)} = \frac{\boldsymbol{\sigma}_{(i)}}{(\boldsymbol{\sigma}_{(i)}^{\mathrm{T}}\boldsymbol{\sigma}_{(i)}+1)^{\frac{1}{4}}}$ and $\boldsymbol{\sigma}_{nb(i)} = \frac{\boldsymbol{\sigma}_{(i)}}{\sqrt{\boldsymbol{\sigma}_{(i)}^{\mathrm{T}}\boldsymbol{\sigma}_{(i)}+1}}$.

Let $\boldsymbol{u}_\epsilon = a\boldsymbol{D}_1\boldsymbol{\sigma}_{nb(i)}\epsilon_1$. If $\epsilon_1 = 0$, then $\boldsymbol{u}_\epsilon = 0$, and (25) reduces to

$$\dot{\tilde{\boldsymbol{W}}}_{(i)} = -a\boldsymbol{D}_1\boldsymbol{\sigma}_{na(i)}\boldsymbol{\sigma}_{na(i)}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{(i)}. \tag{26}$$

Denote the equilibrium of system (25) by $\boldsymbol{c}_{\tilde{W}}$. Under Assumption 4, $\boldsymbol{\sigma}_{na(i)}$ is PE. Under Lemma 1, the origin (i.e. $\boldsymbol{c}_{\tilde{W}} = 0$) of the system (26) is exponentially stable. That is, $\tilde{\boldsymbol{W}}_{(i)}$ converges to zero exponentially.

In the case of $\boldsymbol{u}_\epsilon \neq 0$, it is straightforward to show that (25) has non-zero equilibrium (i.e. $\boldsymbol{c}_{\tilde{W}} \neq 0$), and that $\tilde{\boldsymbol{W}}_{(i)}$ converges to $\boldsymbol{c}_{\tilde{W}}$ exponentially. Since $\|\boldsymbol{\sigma}_{nb(i)}\| < 1$ and $\epsilon_1 \leq b_{\epsilon_1}$, we have $\|\boldsymbol{u}_\epsilon\| \leq b_{u_\epsilon}$ for $b_{u_\epsilon} \in \mathbb{R}^+$ that can be arbitrarily small given sufficient number of suitable activation functions being provided. Therefore, there exists a bound

$b_{\tilde{W}} \in \mathbb{R}^+$ such that $\|\boldsymbol{c}_{\tilde{W}}\| \leq b_{\tilde{W}}$, and $b_{\tilde{W}} \to 0$ with the number of activation functions $n_n \to \infty$. □

Remark 9. As can be seen from the proof of Theorem 1, the MVFA has no direct influence on critic NN weights convergence when considering the 'Policy Evaluation' step only. Exponential stability is primarily due to the admissible control being evaluated. However, the overall system stability in the case of complete synchronous policy iteration (SPI) needs to be further analysed, where the control policy is replaced by a dynamically varying approximation.

4.2. Synchronous policy iteration

As discussed in Remark 7, instability may result when directly implementing the approximated control policy (20) for complete SPI. In this subsection, closed-loop stability under the proposed alternative single-critic scheme with the MVFA is investigated.

Definition 3 (Uniformly Ultimately Bounded)**.** The states $\boldsymbol{x}(t)$ of a dynamic system with initial states $\boldsymbol{x}_0 \triangleq \boldsymbol{x}(t_0)$ is regarded as uniformly ultimately bounded (UUB) about equilibrium $\boldsymbol{x}_e \in \mathbb{R}^{n_x}$ if there exist a compact set $\boldsymbol{\Omega} \in \mathbb{R}^{n_x}$, a finite constant $b_e \in \mathbb{R}^+$ and a time $t_d(b_e, \boldsymbol{x}_0) \in \mathbb{R}^+$ such that $\|\boldsymbol{x}(t) - \boldsymbol{x}_e\| \leq b_e$ for any $\boldsymbol{x}_0 \in \boldsymbol{\Omega}$ and $t \geq t_0 + t_d$ (Lewis, Jagannathan, & Yesildirek, 1999).

Theorem 2. *Consider a system as in (1). Let (9) approximate its value function, with the control policy given by (20). Under Assumptions 1 to 4 and the online tuning law (22), the states $\boldsymbol{x}$ of the system as well as the the critic NN weights estimation error $\tilde{\boldsymbol{W}} = \boldsymbol{W}^* - \hat{\boldsymbol{W}}$ remain UUB during online tuning, if the parameter matrix $\boldsymbol{P}$ in (9) is selected to satisfy $\|\boldsymbol{P}\| > m_P$, for a finite scalar $m_P \in \mathbb{R}^+$.*

Proof. Consider

$$\begin{aligned} L &= \hat{V} + \frac{1}{2}\tilde{\boldsymbol{W}}^{\mathrm{T}}(\kappa_2 \boldsymbol{D}_1)^{-1}\tilde{\boldsymbol{W}} \\ &= L_v + L_w, \end{aligned} \tag{27}$$

where $L_v = \hat{V}$ and $L_w = \frac{1}{2}\tilde{\boldsymbol{W}}^{\mathrm{T}}(\kappa_2 \boldsymbol{D}_1)^{-1}\tilde{\boldsymbol{W}}$.

With (1), (19) and (20), there is

$$\begin{aligned} \dot{L}_v &= (\boldsymbol{f} + \boldsymbol{g}\hat{\boldsymbol{u}})^{\mathrm{T}}\nabla\hat{V} \\ &= (\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{W}})^{\mathrm{T}}\left[\boldsymbol{f} - \frac{1}{2}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{W}})\right] \\ &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{f} + \hat{\boldsymbol{W}}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{f} - \frac{1}{2}\hat{\boldsymbol{W}}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{W}} \\ &\quad - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{W}} - \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x}. \end{aligned} \tag{28}$$

Let $\boldsymbol{G} = \boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}$. With $\hat{\boldsymbol{W}} = \boldsymbol{W}^* - \tilde{\boldsymbol{W}}$, equation (28) becomes

$$\begin{aligned} \dot{L}_v &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{f} + \boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{f} - \tilde{\boldsymbol{W}}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{f} \\ &\quad - \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^* \end{aligned}$$

$$+ \boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}} - \frac{1}{2}\tilde{\boldsymbol{W}}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}}$$
$$+ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}} - \frac{1}{2}\boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}. \tag{29}$$

In regard to the second term in (27), considering (22), we have

$$\begin{aligned}\dot{L}_w &= \tilde{\boldsymbol{W}}^{\mathrm{T}}(\kappa_2\boldsymbol{D}_1)^{-1}\dot{\tilde{\boldsymbol{W}}} = -\tilde{\boldsymbol{W}}^{\mathrm{T}}(\kappa_2\boldsymbol{D}_1)^{-1}\dot{\hat{\boldsymbol{W}}} \\ &= \tilde{\boldsymbol{W}}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}(\boldsymbol{f}+\boldsymbol{g}\hat{\boldsymbol{u}})e_2.\end{aligned} \tag{30}$$

By comparing (4) and (20), there is

$$\hat{\boldsymbol{u}} = \boldsymbol{u}^{*} + \frac{1}{2}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}(\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}} + \nabla\varepsilon). \tag{31}$$

Let $\boldsymbol{z} = \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}}$. Rewriting (30) using (31) gives

$$\begin{aligned}\dot{L}_w &= \frac{1}{2}\tilde{\boldsymbol{W}}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\varepsilon e_2 + \tilde{\boldsymbol{W}}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})e_2 \\ &\quad + \frac{1}{2}\tilde{\boldsymbol{W}}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}}e_2 \\ &= \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})e_2 + \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon e_2 + \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}e_2.\end{aligned} \tag{32}$$

Subtracting (21) from (5) yields

$$\begin{aligned}e_2 &= (\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{W}})^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\hat{\boldsymbol{u}}) + \hat{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{R}\hat{\boldsymbol{u}} - \boldsymbol{u}^{*\mathrm{T}}\boldsymbol{R}\boldsymbol{u}^{*} \\ &\quad - (\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*} + \nabla\varepsilon)^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*}).\end{aligned} \tag{33}$$

By using (4), (10), (19) and (20), the individual terms in (33) have expressions of

$$\begin{aligned}&(\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{W}})^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\hat{\boldsymbol{u}}) \\ &= (\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})^{\mathrm{T}}(\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*} - \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}}) \\ &\quad + \frac{1}{2}(\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}(\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}} + \nabla\varepsilon) \\ &\quad - \frac{1}{2}(\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}(\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}} + \nabla\varepsilon),\end{aligned} \tag{34}$$

$$\begin{aligned}&\hat{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{R}\hat{\boldsymbol{u}} \\ &= \frac{1}{4}(\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}) \\ &\quad - \frac{1}{2}(\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}} \\ &\quad + \frac{1}{4}(\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}},\end{aligned} \tag{35}$$

$$
\begin{aligned}
&\boldsymbol{u}^{*\mathrm{T}}\boldsymbol{R}\boldsymbol{u}^{*} \\
=&\frac{1}{4}(\boldsymbol{P}\boldsymbol{x}+\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{P}\boldsymbol{x}+\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}) \\
&+\frac{1}{2}(\boldsymbol{P}\boldsymbol{x}+\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\varepsilon \\
&+\frac{1}{4}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\varepsilon. \qquad (36)
\end{aligned}
$$

Substituting (34), (35), and (36) back into (33) gives

$$
\begin{aligned}
e_2 =&-(\nabla\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}})^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})-\nabla\varepsilon^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*}) \\
&-\frac{1}{2}(\nabla\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\varepsilon-\frac{1}{4}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\varepsilon \\
&-\frac{1}{4}(\nabla\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}})^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}}. \qquad (37)
\end{aligned}
$$

It can be seen from (37) that for a given set of NN hidden-layer neurons of a finite number n_n, the minimum of e_2, denoted by ϵ_2, is reached when $\tilde{\boldsymbol{W}}=0$:

$$
\epsilon_2=-\nabla\varepsilon^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})-\frac{1}{4}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}\nabla\varepsilon. \qquad (38)
$$

Under Assumptions 1 and 3, $\nabla\varepsilon$ and $(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})$ are bounded. Thus, there exist a finite constant $b_{\epsilon_2}\in\mathbb{R}^{+}$ such that $\epsilon_2\leq b_{\epsilon_2}$. Since $\varepsilon\to 0$ and $\nabla\varepsilon\to 0$ as the number of suitable activation functions n_n increases infinitely (Finlayson, 1972), it is straightforward to show that $\epsilon_2\to 0,\ \forall\boldsymbol{x}\neq 0$, if $n_n\to\infty$. As a special case, $\epsilon_2=0$ if $\nabla\varepsilon=0,\ \forall\boldsymbol{x}\neq 0$.

Substituting (33) for e_2 in (32) yields

$$
\begin{aligned}
\dot{L}_w =&-\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})^{\mathrm{T}}\boldsymbol{z}-\frac{3}{8}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon \\
&-\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})^{\mathrm{T}}\nabla\varepsilon-\frac{1}{8}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} \\
&-\frac{3}{4}\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\nabla\varepsilon^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*}) \\
&-\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon-\frac{1}{8}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon \\
&-\frac{1}{4}\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon-\frac{1}{8}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon \\
&-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\nabla\varepsilon^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})-\frac{1}{4}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}. \qquad (39)
\end{aligned}
$$

Note that the first term in (39) can be expanded as

$$
\begin{aligned}
&-\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})(\boldsymbol{f}+\boldsymbol{g}\boldsymbol{u}^{*})^{\mathrm{T}}\boldsymbol{z} \\
=&-\boldsymbol{z}^{\mathrm{T}}\boldsymbol{f}\boldsymbol{f}^{\mathrm{T}}\boldsymbol{z}-\frac{1}{4}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}\boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{z} \\
&-\frac{1}{4}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}-\frac{1}{4}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}
\end{aligned}
$$

$$
\begin{aligned}
&+ \boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\boldsymbol{f}^{\mathrm{T}}\boldsymbol{z} - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} \\
&+ \boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}\boldsymbol{f}^{\mathrm{T}}\boldsymbol{z} - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} \\
&+ \boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\boldsymbol{f}^{\mathrm{T}}\boldsymbol{z} - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{z}. \qquad (40)
\end{aligned}
$$

Combining (29) and (39) gives:

$$
\dot{L} = T_1 + T_2 + T_3 + T_4 + T_5, \qquad (41)
$$

where

$$
\begin{aligned}
T_1 = &- \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{f} + \boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{f} \\
&- \boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}, \qquad (42)
\end{aligned}
$$

$$
\begin{aligned}
T_2 = &- \frac{1}{4}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{z} \\
&- \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} \\
&+ \boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\boldsymbol{f}^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}\boldsymbol{f}^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\boldsymbol{f}^{\mathrm{T}}\boldsymbol{z} \\
&- \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^{*})\boldsymbol{z}\boldsymbol{G}\nabla\varepsilon - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\nabla\varepsilon^{\mathrm{T}}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^{*}) \\
&- \boldsymbol{z}^{\mathrm{T}}\boldsymbol{f} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} + \boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{z}, \qquad (43)
\end{aligned}
$$

$$
\begin{aligned}
T_3 = &- \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} - \frac{1}{4}\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^{*})\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon \\
&- \frac{1}{8}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\nabla\varepsilon^{\mathrm{T}}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^{*}) \\
&- \boldsymbol{z}^{\mathrm{T}}\boldsymbol{f}\boldsymbol{f}^{\mathrm{T}}\boldsymbol{z} - \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^{*})(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^{*})^{\mathrm{T}}\nabla\varepsilon \\
&- \frac{1}{4}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*}\boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{z} \\
&- \frac{1}{8}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon\nabla\varepsilon^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon, \qquad (44)
\end{aligned}
$$

$$
\begin{aligned}
T_4 = &- \frac{1}{8}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} - \frac{3}{4}\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^{*})\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} \\
&- \frac{3}{8}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\nabla\varepsilon, \qquad (45)
\end{aligned}
$$

$$
T_5 = -\frac{1}{2}\boldsymbol{W}^{*\mathrm{T}}\bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{W}^{*} - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}. \qquad (46)
$$

Now introduce bounds to (42).

As $\boldsymbol{G} = \boldsymbol{g}\boldsymbol{R}^{-1}\boldsymbol{g}^{\mathrm{T}}$, the rank of $\boldsymbol{G}$ is

$$\operatorname{rank}(\boldsymbol{G}) = \operatorname{rank}(\boldsymbol{g}) < n_x. \tag{47}$$

It follows that there exist kernel

$$\ker(\boldsymbol{G}\boldsymbol{P}\boldsymbol{r}) = \{\boldsymbol{r} \in \mathbb{R}^{n_x} \mid \boldsymbol{G}\boldsymbol{P}\boldsymbol{r} = 0\}. \tag{48}$$

For nonlinear systems as in (1), since $\boldsymbol{x}$ and $\boldsymbol{z}$ are explicitly governed by (1) instead of being random, the existence of $\boldsymbol{x} = \ker(\boldsymbol{G}\boldsymbol{P}\boldsymbol{r})$ and corresponding effects to the system is rendered negligible. Accordingly, we focus on $\boldsymbol{x} \neq \ker(\boldsymbol{G}\boldsymbol{P}\boldsymbol{r})$ in this paper. In this case, $\boldsymbol{G}$ is positive-definite and symmetric, and under Assumptions 2 and 3, there is

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x} \geq m_1\|\boldsymbol{P}\|^2\|\boldsymbol{x}\|^2, \tag{49}$$

where constant $m_1 \in \mathbb{R}^+$. Also, there is $\|\boldsymbol{G}\| \leq b_G$ for constant $b_G \in \mathbb{R}^+$.

Together with Assumption 3, the following inequality holds:

$$\begin{aligned} T_1 \leq & \left(-\frac{1}{2}m_1\|\boldsymbol{P}\|^2 + b_f\|\boldsymbol{P}\| + b_G b_\Phi \|\boldsymbol{W}^*\|\|\boldsymbol{P}\| + b_\Phi b_f\|\boldsymbol{W}^*\|\right)\|\boldsymbol{x}\|^2 \\ = & -\frac{1}{2}m_1\|\boldsymbol{x}\|^2\left(\|\boldsymbol{P}\|^2 - c_1\|\boldsymbol{P}\| - c_2\right), \end{aligned} \tag{50}$$

where

$$\begin{aligned} c_1 &= \frac{2\left(b_f + b_G b_\Phi\|\boldsymbol{W}^*\|\right)}{m_1}, \\ c_2 &= \frac{2b_\Phi b_f\|\boldsymbol{W}^*\|}{m_1}. \end{aligned}$$

Under Assumption 4, if $\|\boldsymbol{P}\|^2 - c_1\|\boldsymbol{P}\| - c_2 \geq 0$, then $T_1 \leq 0$. This requires

$$\|\boldsymbol{P}\| \geq \frac{c_1 + \sqrt{c_1^2 + 4c_2}}{2} \triangleq p_1. \tag{51}$$

In T_2, similarly to the case of T_1, we consider the circumstances of $\boldsymbol{x} \neq \ker(\boldsymbol{G}\boldsymbol{P}\boldsymbol{r})$ and $\boldsymbol{z} \neq \ker(\boldsymbol{G}\boldsymbol{r})$. Then there is a finite constant $m_2 \in \mathbb{R}^+$ such that

$$\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} \geq m_2\|\boldsymbol{P}\|^2\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|^2. \tag{52}$$

Given Assumption 3, we have $\|\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^*\| \leq b_{\dot{x}}\|\boldsymbol{x}\|$, for a finite constant $b_{\dot{x}} \in \mathbb{R}^+$. Hence,

$$\begin{aligned} T_2 \leq & \left(-\frac{1}{4}m_2\|\boldsymbol{P}\|^2 + \frac{1}{2}b_G^2 b_\varepsilon\|\boldsymbol{P}\| + b_G b_f\|\boldsymbol{P}\| + \frac{1}{2}b_G^2 b_\Phi\|\boldsymbol{W}^*\|\|\boldsymbol{P}\|\right. \\ & \left. + \frac{1}{2}b_G^2 b_\Phi b_\varepsilon\|\boldsymbol{W}^*\| + b_G b_\varepsilon b_f + b_G b_\Phi b_f\|\boldsymbol{W}^*\| + \frac{3}{2}b_{\dot{x}} b_G b_\varepsilon\right)\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|^2 \\ & + \left(b_G\|\boldsymbol{P}\| + b_f + b_G b_\Phi\|\boldsymbol{W}^*\|\right)\|\boldsymbol{x}\|\,\|\boldsymbol{z}\| \\ = & -\frac{1}{4}m_2\left(\|\boldsymbol{P}\|^2 - d_1\|\boldsymbol{P}\| - d_2\right)\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|^2 \end{aligned}$$

$$+ \left(b_G \|\boldsymbol{P}\| + b_f + b_G b_\Phi \|\boldsymbol{W}^*\|\right) \|\boldsymbol{x}\| \, \|\boldsymbol{z}\|, \tag{53}$$

where

$$d_1 = \frac{2b_G^2 b_\Phi \|\boldsymbol{W}^*\| + 2b_G^2 b_\varepsilon + 4b_G b_f}{m_2},$$
$$d_2 = \frac{(2b_G^2 b_\Phi b_\varepsilon + 4b_G b_\Phi b_f)\|\boldsymbol{W}^*\| + 4b_G b_\varepsilon b_f + 6b_{\dot{x}} b_G b_\varepsilon}{m_2}.$$

Let

$$c_3 = \|\boldsymbol{P}\|^2 - d_1 \|\boldsymbol{P}\| - d_2,$$
$$c_4 = b_G \|\boldsymbol{P}\| + b_f + b_G b_\Phi \|\boldsymbol{W}^*\|.$$

Then (53) can be rewritten as

$$T_2 \leq -\frac{1}{4} m_2 c_3 \left(\|\boldsymbol{x}\| \, \|\boldsymbol{z}\| - \frac{4c_4}{m_2 c_3}\right) \|\boldsymbol{x}\| \, \|\boldsymbol{z}\|. \tag{54}$$

It is clear that $T_2 \leq 0$ if $c_3 \left(\|\boldsymbol{x}\| \, \|\boldsymbol{z}\| - \frac{4c_4}{m_2 c_3}\right) \geq 0$. That is, if

$$\|\boldsymbol{P}\| > \frac{d_1 + \sqrt{d_1^2 + 4d_2}}{2} \triangleq p_2, \tag{55}$$

then $c_3 > 0$, and in this case, $T_2 \leq 0$ whenever

$$\|\boldsymbol{x}\| \, \|\boldsymbol{z}\| \geq \frac{4c_4}{m_2 c_3}. \tag{56}$$

Regarding T_3, for cases of $\boldsymbol{z} \neq \ker(\boldsymbol{G}\boldsymbol{r})$ and $\nabla\varepsilon \neq \ker(\boldsymbol{G}\boldsymbol{r})$, there exist constants $m_3,\ m_4,\ m_5,\ m_6 \in \mathbb{R}^+$ such that the following inequalities hold:

$$\boldsymbol{z}^{\mathrm{T}} \boldsymbol{G} \nabla\varepsilon \nabla\varepsilon^{\mathrm{T}} \boldsymbol{G} \boldsymbol{z} \geq m_3 \|\boldsymbol{x}\|^4 \left\|\tilde{\boldsymbol{W}}\right\|^2, \tag{57}$$

$$\boldsymbol{z}^{\mathrm{T}} \boldsymbol{G} \boldsymbol{z} \nabla\varepsilon^{\mathrm{T}} \boldsymbol{G} \nabla\varepsilon \geq m_4 \|\boldsymbol{x}\|^4 \left\|\tilde{\boldsymbol{W}}\right\|^2, \tag{58}$$

$$\boldsymbol{z}^{\mathrm{T}} \boldsymbol{f} \boldsymbol{f}^{\mathrm{T}} \boldsymbol{z} \geq m_5 \|\boldsymbol{x}\|^4 \left\|\tilde{\boldsymbol{W}}\right\|^2, \tag{59}$$

$$\boldsymbol{z}^{\mathrm{T}} \boldsymbol{G} \bar{\nabla} \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* \boldsymbol{W}^{*\mathrm{T}} \bar{\nabla} \boldsymbol{\Phi} \boldsymbol{G} \boldsymbol{z} \geq m_6 \|\boldsymbol{x}\|^4 \left\|\tilde{\boldsymbol{W}}\right\|^2. \tag{60}$$

Therefore, we have from (44) that

$$\begin{aligned} T_3 \leq & \left(-\frac{1}{2} m_3 \left\|\tilde{\boldsymbol{W}}\right\| - \frac{1}{8} m_4 \left\|\tilde{\boldsymbol{W}}\right\| - m_5 \left\|\tilde{\boldsymbol{W}}\right\| - \frac{1}{4} m_6 \left\|\tilde{\boldsymbol{W}}\right\| \right. \\ & \left. + \frac{1}{4} b_{\dot{x}} b_\varepsilon^2 b_G b_\Phi + \frac{1}{2} b_{\dot{x}} b_\varepsilon^2 b_G b_\Phi + b_{\dot{x}}^2 b_\varepsilon b_\Phi + \frac{1}{8} b_\varepsilon^3 b_G^2 b_\Phi \right) \left\|\tilde{\boldsymbol{W}}\right\| \|\boldsymbol{x}\|^4 \\ = & - d_3 \|\boldsymbol{x}\|^4 \left\|\tilde{\boldsymbol{W}}\right\| \left(\left\|\tilde{\boldsymbol{W}}\right\| - \frac{d_4}{d_3}\right), \end{aligned} \tag{61}$$

where

$$d_3 = \frac{1}{2}m_3 + \frac{1}{8}m_4 + m_5 + \frac{1}{4}m_6,$$
$$d_4 = \frac{3}{4}b_{\dot{x}}b_{\varepsilon}^2 b_G b_\Phi + b_{\dot{x}}^2 b_\varepsilon b_\Phi + \frac{1}{8}b_\varepsilon^3 b_G^2 b_\Phi.$$

As a result, under Assumption 4, $T_3 \leq 0$ when

$$\left\|\tilde{\boldsymbol{W}}\right\| \geq \frac{d_4}{d_3}. \tag{62}$$

With regard to T_4, for $\boldsymbol{z} \neq \ker(\boldsymbol{G}\boldsymbol{r})$, there is

$$\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{z} \geq m_7\|\boldsymbol{x}\|^4\left\|\tilde{\boldsymbol{W}}\right\|^4 \tag{63}$$

for a constant $m_7 \in \mathbb{R}^+$. Then from (45),

$$\begin{aligned} T_4 \leq & -\frac{1}{8}m_7\|\boldsymbol{x}\|^4\left\|\tilde{\boldsymbol{W}}\right\|^4 + \frac{3}{4}b_{\dot{x}}b_G b_\Phi^3\|\boldsymbol{x}\|^4\left\|\tilde{\boldsymbol{W}}\right\|^3 \\ & + \frac{3}{8}b_\varepsilon b_G^2 b_\Phi^3\|\boldsymbol{x}\|^4\left\|\tilde{\boldsymbol{W}}\right\|^3 \\ = & -\frac{1}{8}m_7\left(\left\|\tilde{\boldsymbol{W}}\right\| - \frac{6b_{\dot{x}}b_G b_\Phi^3 + 3b_\varepsilon b_G^2 b_\Phi^3}{m_7}\right)\|\boldsymbol{x}\|^4\left\|\tilde{\boldsymbol{W}}\right\|^3. \end{aligned} \tag{64}$$

Therefore, under Assumption 4, $T_4 \leq 0$ requires

$$\left\|\tilde{\boldsymbol{W}}\right\| \geq \frac{6b_{\dot{x}}b_G b_\Phi^3 + 3b_\varepsilon b_G^2 b_\Phi^3}{m_7}. \tag{65}$$

It is easy to see from (46) that $T_5 \leq 0$. Thus, it can be concluded from (51), (55), (56), (62), and (65) that (41) is negative if

$$\|\boldsymbol{P}\| > \max(p_1, p_2) \triangleq m_P, \tag{66}$$
$$\|\boldsymbol{x}\|\,\|\boldsymbol{z}\| > \frac{4c_4}{m_2 c_3}, \tag{67}$$
$$\left\|\tilde{\boldsymbol{W}}\right\| > \max\left(\frac{d_4}{d_3},\ \frac{6b_{\dot{x}}b_G b_\Phi^3 + 3b_\varepsilon b_G^2 b_\Phi^3}{m_7}\right). \tag{68}$$

Since $\boldsymbol{z} = \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}}\tilde{\boldsymbol{W}}$, and $\boldsymbol{x}$ is PE under Assumption 4, equation (67) also establishes a bound for $\|\tilde{\boldsymbol{W}}\|$. Thus, equations (67) and (68) together, show that $\tilde{\boldsymbol{W}}$ is UUB. Finally, upon satisfaction of (66), (67), and (68), the UUB stability holds for the system states and NN weights estimation errors. □

The following theorem further reveals the link between the UUB boundedness of NN weight estimation errors and the system states stability.

Theorem 3. *Given Assumptions 1 to 4 and the tuning law provided by (22), the nonlinear system as in (1) remains asymptotically stable in the online learning process*

under the control given by (20), if matrix $\boldsymbol{P}$ *in (18) satisfies* $\|\boldsymbol{P}\| > b_{mP}$ *for a scalar* $b_{mP} \in \mathbb{R}^+$.

Proof. The Lyapunov function candidate is selected to be $L_V = \hat{V}$, the time derivative of which is:

$$
\begin{aligned}
\dot{L}_V &= (\boldsymbol{f} + \boldsymbol{g}\hat{\boldsymbol{u}})^{\mathrm{T}} \left(\frac{\partial \hat{V}}{\partial \boldsymbol{x}} \right) \\
&= (\boldsymbol{f} + \boldsymbol{g}\hat{\boldsymbol{u}})^{\mathrm{T}} (\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \hat{\boldsymbol{W}}) \\
&= (\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^* - \boldsymbol{g}\tilde{\boldsymbol{u}})^{\mathrm{T}} (\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* - \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \tilde{\boldsymbol{W}}) \\
&= (\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^*)^{\mathrm{T}} \left(\frac{\partial V^*}{\partial \boldsymbol{x}} \right) - (\boldsymbol{f} + \boldsymbol{g}\boldsymbol{u}^*)^{\mathrm{T}} \nabla\varepsilon \\
&\quad + \frac{1}{2} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{G} (\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \tilde{\boldsymbol{W}} + \nabla\varepsilon) \\
&\quad + \frac{1}{2} \boldsymbol{W}^{*\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} (\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \tilde{\boldsymbol{W}} + \nabla\varepsilon) - \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi}\boldsymbol{f} \\
&\quad + \frac{1}{2} \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} (\boldsymbol{P}\boldsymbol{x} + \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* + \nabla\varepsilon) \\
&\quad - \frac{1}{2} \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} (\bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \tilde{\boldsymbol{W}} + \nabla\varepsilon) \\
&= -\bar{Q} - \boldsymbol{u}^{*\mathrm{T}} \boldsymbol{R}\boldsymbol{u}^* - \nabla\varepsilon^{\mathrm{T}} \boldsymbol{f} - \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi}\boldsymbol{f} + \nabla\varepsilon^{\mathrm{T}} \boldsymbol{G}\boldsymbol{P}\boldsymbol{x} \\
&\quad + \frac{1}{2} \nabla\varepsilon^{\mathrm{T}} \boldsymbol{G} \nabla\varepsilon + \nabla\varepsilon^{\mathrm{T}} \boldsymbol{G} \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* + \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x} \\
&\quad + \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* - \frac{1}{2} \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \tilde{\boldsymbol{W}} \\
&= -\bar{Q} - \frac{1}{4} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{G}\boldsymbol{P}\boldsymbol{x} - \frac{1}{2} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{G} \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* - \nabla\varepsilon^{\mathrm{T}} \boldsymbol{f} \\
&\quad + \frac{1}{2} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{G} \nabla\varepsilon - \frac{1}{4} \boldsymbol{W}^{*\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* + \frac{1}{4} \nabla\varepsilon^{\mathrm{T}} \boldsymbol{G} \nabla\varepsilon \\
&\quad + \frac{1}{2} \boldsymbol{W}^{*\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} \nabla\varepsilon - \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi}\boldsymbol{f} + \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{P}\boldsymbol{x} \\
&\quad + \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* - \frac{1}{2} \tilde{\boldsymbol{W}}^{\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \tilde{\boldsymbol{W}}.
\end{aligned} \tag{69}
$$

Since $\bar{Q}(\boldsymbol{x}) > 0$, there exists constant $b_Q \in \mathbb{R}^+$ such that $b_Q \|\boldsymbol{x}\|^2 \leq \bar{Q}(x)$. Given that $\boldsymbol{x}$ is explicitly governed by system (1), the case of $\boldsymbol{x} \neq \ker(\boldsymbol{G})$ is considered. It is straightforward to see that there exist constants $b_{GL} \in \mathbb{R}^+$ and $b_{m4} \in \mathbb{R}^+$ such that

$$
b_{GL} \leq \|\boldsymbol{G}\|,
$$

and

$$
b_{m4} \|\boldsymbol{x}\|^2 \leq \left\| \boldsymbol{W}^{*\mathrm{T}} \bar{\nabla}\boldsymbol{\Phi} \boldsymbol{G} \bar{\nabla}\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{W}^* \right\|.
$$

Following the results of Theorem 2, it is known that $0 \leq \|\tilde{\boldsymbol{W}}\| \leq b_{\tilde{W}}$. Then (69)

can be upper bounded as:

$$\begin{aligned}\dot{L}_V \leq (&-b_Q - \frac{1}{4}b_{GL}\|\boldsymbol{P}\|^2 + \frac{1}{2}b_{GU}b_\varepsilon\|\boldsymbol{P}\| \\ &+ \frac{1}{2}b_{GU}b_\phi\|\boldsymbol{W}^*\|\|\boldsymbol{P}\| - \frac{1}{4}b_{m4} + b_\varepsilon b_f \\ &+ \frac{1}{2}b_\phi b_{GU}b_\varepsilon\|\boldsymbol{W}^*\| + \frac{1}{4}b_\varepsilon^2 b_{GU} + b_\phi b_f b_{\tilde{W}} \\ &+ b_\phi b_{GU}b_{\tilde{W}}\|\boldsymbol{P}\| + b_\phi^2 b_{GU}b_{\tilde{W}}\|\boldsymbol{W}^*\|)\|\boldsymbol{x}\|^2 \\ = &-\|\boldsymbol{x}\|^2\left(\frac{1}{4}b_{GL}\|\boldsymbol{P}\|^2 - \eta_1\|\boldsymbol{P}\| - \eta_2\right). \end{aligned} \tag{70}$$

where

$$\eta_1 = \frac{1}{2}b_{GU}b_\phi\|\boldsymbol{W}^*\| + \frac{1}{2}b_{GU}b_\varepsilon + b_\phi b_{GU}b_{\tilde{W}},$$

and

$$\begin{aligned}\eta_2 = &\frac{1}{2}b_\phi b_{GU}b_\varepsilon\|\boldsymbol{W}^*\| + \frac{1}{4}b_\varepsilon^2 b_{GU} + b_\varepsilon b_f + b_\phi b_f b_{\tilde{W}} \\ &+ b_\phi^2 b_{GU}b_{\tilde{W}}\|\boldsymbol{W}^*\| - b_Q - \frac{1}{4}b_{m4}. \end{aligned} \tag{71}$$

Equation (70) shows that $\dot{L}_V$ is negative, and thus $\|\boldsymbol{x}\|$ is bounded, as long as

$$\|\boldsymbol{P}\| \geq \frac{2\eta_1 + 2\sqrt{\eta_1^2 + b_{GL}\eta_2}}{b_{GL}} \triangleq b_{mP}. \tag{72}$$

It follows that $\ddot{L}_V = \frac{d\dot{L}_V}{dt}$ is a function of $\boldsymbol{x}$ and $\tilde{\boldsymbol{W}}$, and $\ddot{L}_V$ is also bounded as $\|\boldsymbol{x}\|$ and $\|\tilde{\boldsymbol{W}}\|$ are bounded. As a result, asymptotic stability applies to the system states $\boldsymbol{x}$. □

Remark 10. As can be seen from Theorems 2 and 3, the proposed MVFA establishes a direct link to closed-loop stability. With the MVFA, no special stabilising tuning laws are required for the NNs in critic and actor, and during online learning the SPI under the resulted single-critic configuration remains stable with simple gradient descent tuning.

5. Numerical studies

This section presents two simulation examples. Finding the optimal control law for a nonlinear model with a known value function is first introduced to verify the proposed method. The second example then demonstrates the use of the proposed controller in a practical engineering application where a nonlinear system with higher dimension is involved.

5.1. Nonlinear example

The following nonlinear system is considered (Vamvoudakis & Lewis, 2010), with

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2\left\{1 - [\cos(2x_1) + 2]^2\right\} \end{bmatrix},$$

and

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}.$$

For $\boldsymbol{Q} = \boldsymbol{I}_{2\times 2}$ and $R = 1$, the corresponding V^* and $\boldsymbol{u}^*$ are known to be

$$V^* = \frac{1}{2}x_1^2 + x_2^2, \tag{73}$$

and

$$u^* = -\left[\cos(2x_1) + 2\right] x_2, \tag{74}$$

respectively, as given in Vamvoudakis and Lewis (2010).

The critic NN has activation functions of

$$\boldsymbol{\Phi} = [x_1^2,\ x_1x_2,\ x_2^2]^{\mathrm{T}},$$

with NN weights being

$$\hat{\boldsymbol{W}} = [\hat{W}_1,\ \hat{W}_2,\ \hat{W}_3]^{\mathrm{T}}.$$

In simulation, $\boldsymbol{P} = 10\boldsymbol{I}_{2\times 2}$, $a = 10$, and $\boldsymbol{D}_3 = [5,\ 1]^{\mathrm{T}}$. System states $\boldsymbol{x}$ and NN weights $\hat{\boldsymbol{W}}$ are initialised to zeros. An exogenous signal

$$\begin{aligned} u_e(t) = 2[&\cos(0.8t) + \sin(t)^2\cos(t) + \sin(2t)^2\cos(0.1t) \\ &+ \sin(-1.2t)^2\cos(0.5t) + \sin(t)^5] \end{aligned}$$

is used to perturb the system for exploration. Note that the total control that enters the process during exploration is the sum of $\hat{u}$ and u_e, which also perturbs the system states $\boldsymbol{x}$ in the meantime. For efficient and effective training with (21) and (22) involved, exploration is implemented in the following manner: The excitation of $u_e(t)$ lasts 0.05 s for every 0.1 s time interval, while the HJB error e_2 in (21) is periodically fed back for calculation during the intervals when $u_e(t)$ is temporarily off (i.e., $e_2 = 0$ if $u_e(t) \neq 0$). $u_e(t)$ is completely turned off at 40 s. 2.2

The trajectories of system states $\boldsymbol{x}$, approximated optimal control $\hat{u}$ and the excitation signal u_e during online learning are plotted in Figures 1 and 2, respectively. Close-up of the excitation signal u_e for the first 2 seconds is shown in Figure 3 for clearer illustration of the special excitation implemented. Weights convergence history of the critic NN is given in Figure 4.

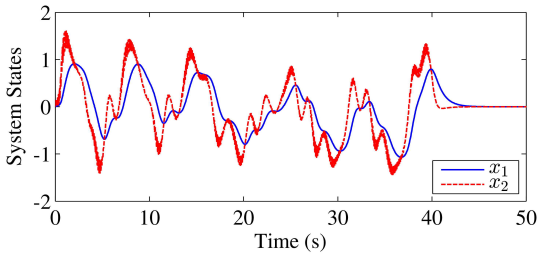

Figure 1. Trajectories of system states during online tuning.

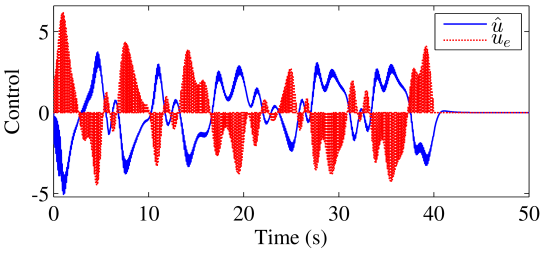

Figure 2. Trajectories of control signals during online tuning.

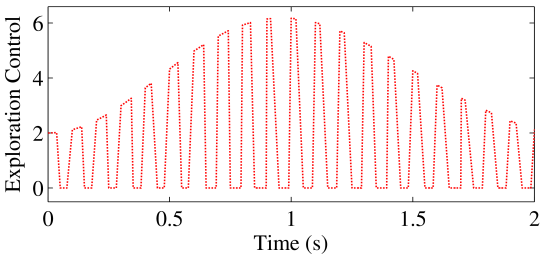

Figure 3. Close-up of excitation signal u_e for the first 2 seconds.

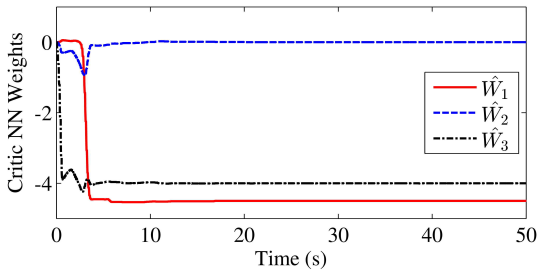

Figure 4. NN parameters convergence during online tuning.

Figure 4 shows that all NN weights settle within 10 s. At the end of training,

$$\hat{\boldsymbol{W}} = [-4.4999,\ -0.0003,\ -3.9996]^{\mathrm{T}}.$$

This yields

$$\hat{V}(x) = 0.5001{x_1}^2 - 0.0003x_1x_2 + 1.0004{x_2}^2$$

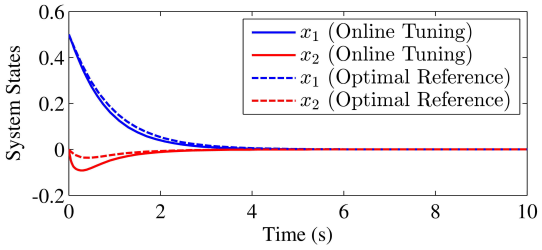

Figure 5. State trajectories of the closed-loop response to the non-zero initial condition under the proposed online tuning scheme (PE unsatisfied in this case) and the known ideal optimal control.

$$\approx \frac{1}{2}x_1^2 + x_2^2,$$

and

$$\begin{aligned}\hat{u}(x) &= -\left[\cos(2x_1) + 2\right](-0.0002x_1 + 1.0004x_2) \\ &\approx -\left[\cos(2x_1) + 2\right]x_2,\end{aligned}$$

which are close approximation to (73) and (74), showing that the convergence of NN weights is reached with good accuracy.

Also note that the approximated optimal control $\hat{u}$ generally mirrors the contour of the excitation signal u_e with slight difference in amplitude. It shows $\hat{u}$ effectively counteracts u_e and maintains closed-loop system states stability during online training. 2.6

In situations when the PE condition may not be satisfied (for example, the closed-loop response is subject to none-zero initial conditions only), $\hat{\boldsymbol{W}}$ may not reach its ideal set $\boldsymbol{W}^*$ as a result. In the following simulations, the convergence of NN parameters and the closed-loop stability is investigated under unsatisfied PE condition. Accordingly, $\boldsymbol{x}(0) = [0.5 \quad 0]^{\mathrm{T}}$ is applied as an initial condition, no probing noise is added, and controller parameters remain the same. The corresponding closed-loop states responses are plotted in Figure 5, and the corresponding control action is given in Figure 6, together with responses under the ideal optimal control supplied for comparison. As can be seen from the figures, states trajectories and control signal under the proposed control scheme are similar to those of the ideal optimal control. The difference in response is due to the approximation error resulted from lack of PE. The NN parameters convergence history is plotted in Figure 7, where the settling value of $\hat{W}_2$ and $\hat{W}_3$ is still far from the ideal one. However, stable closed-loop responses are observed under the proposed algorithm regardless of the differences, as shown by Figures 5 and 6. The cost of the closed-loop response to the none-zero initial condition under the proposed algorithm (i.e., $\hat{V}(\boldsymbol{x}(0))$) together with that under the known ideal optimal control (i.e., $V^*(\boldsymbol{x}(0))$) are evaluated in Figure 8. By recalculating the cost using the continuously updated NN weights, the approximated value function $\hat{V}$ is shown to be converging to the optimal one, in the presence of some approximation error.

5.2. Nonlinear application example

In the following, the proposed controller is used for actively suppressing airfoil flutter. This example demonstrates the capability of the proposed controller in dealing with real-world applications with higher model dimension. A two-degrees-of-freedom

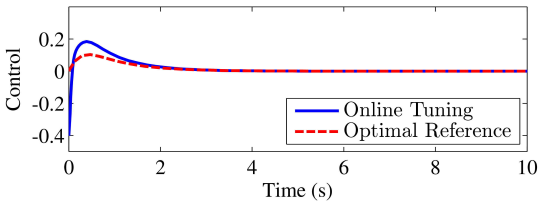

Figure 6. Control input in response to the non-zero initial condition under the proposed online tuning scheme (PE unsatisfied in this case) and the known ideal optimal control.

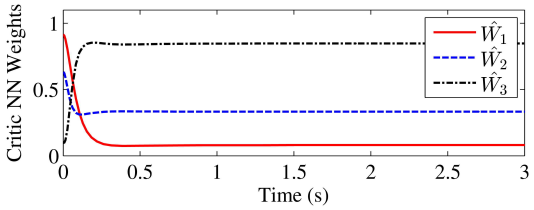

Figure 7. NN parameters convergence history during the closed-loop response to the non-zero initial condition (PE unsatisfied case).

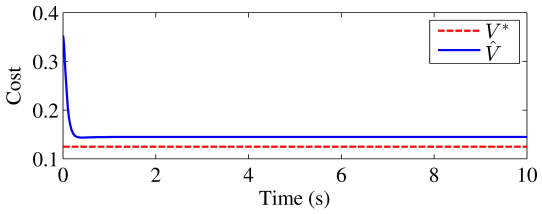

Figure 8. The minimal cost $V^*(\boldsymbol{x}(0))$ of the closed-loop response to the non-zero initial condition and the evolution of the approximated $\hat{V}(\boldsymbol{x}(0))$ (PE unsatisfied case).

(2DOF) nonlinear aeroelastic model in state space with 4 states and 2 control inputs is used (Z. Wang, Behal, & Marzocca, 2011). The aeroelastic model is nonlinear in pitch stiffness, and all parameters are taken from Z. Wang et al. (2011). The critical wind speed is around 11.42m/s, at and above which, the airfoil becomes unstable and starts fluttering when given some excitation or none-zero initial condition.

At the critical wind speed and with the initial condition of $\boldsymbol{x}(0) = [0 \quad 0.1 \quad 0 \quad 0]^{\mathrm{T}}$, the open-loop responses of the airfoil in terms of plunge and pitch motions are plotted in Figure 9.

Given the nonlinear pitch stiffness in polynomial form up to the 3^{rd} order, a total of 65 activation functions up to 4^{th} order are selected for $\boldsymbol{\Phi}(\boldsymbol{x})$. Other parameters are $\boldsymbol{P} = 10\boldsymbol{I}_{4\times4}$, $\boldsymbol{Q} = \boldsymbol{I}_{4\times4}$, $\boldsymbol{R} = \boldsymbol{I}_{2\times2}$, $a = 10$, and $\boldsymbol{D}_3 = [1,\ 1,\ 1,\ 1]^{\mathrm{T}}$. $\hat{\boldsymbol{W}}$ is initialised to zeros. The corresponding control law for each control channel is not listed herein due to space concern.

With the proposed controller turned on, the flutter triggered by the same initial condition can be effectively suppressed, as shown in Figure 10, regardless of whether the controller has been trained or not. However, manifest differences are observed before and after training the proposed controller (using the same techniques as in the previous example), and the post-learning controller shows better performance than the pre-learning one.

To verify that the post-learning controller offers near-optimal control, an offline

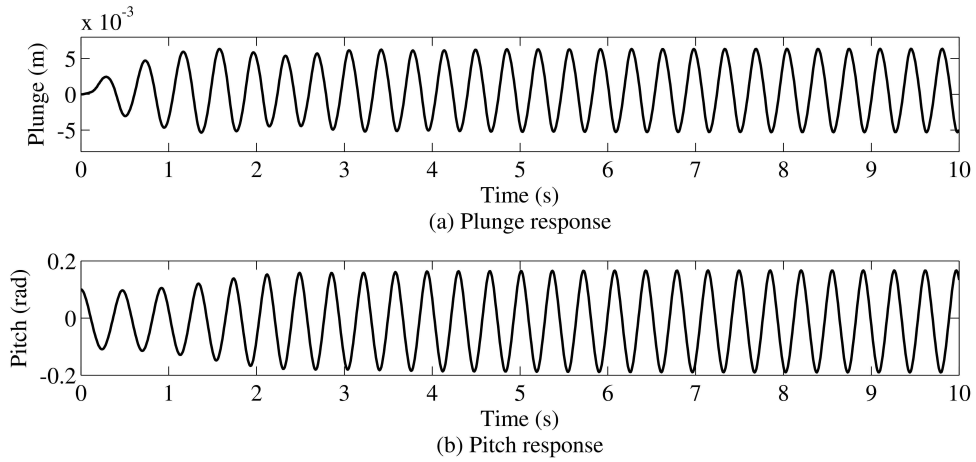

Figure 9. Open-loop response of a 2DOF aeroelastic system at critical wind speed.

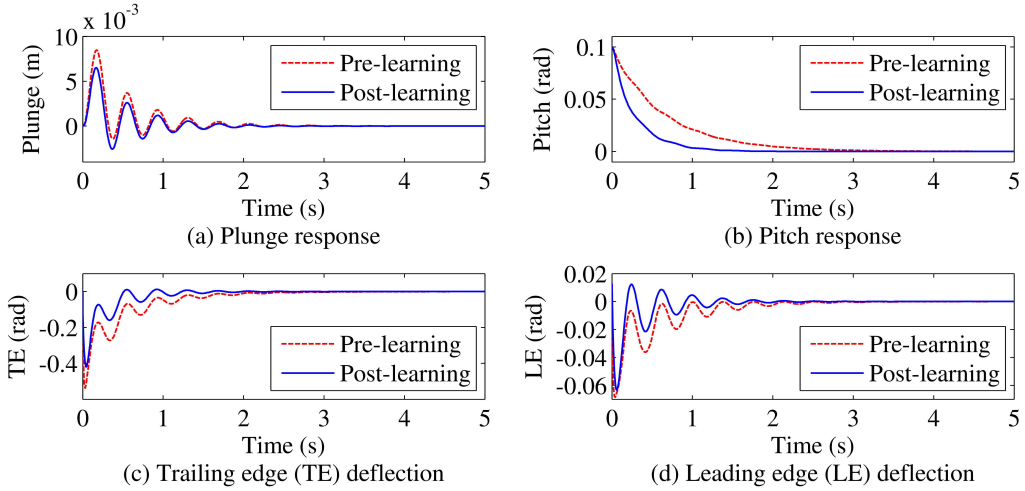

Figure 10. Closed-loop responses of a 2DOF aeroelastic system at critical wind speed with the proposed controller.

method (Abu-Khalaf & Lewis, 2005) is used to synthesise the nonlinear control law for the same conditions as a benchmark. The cost of suppressing flutter triggered by initial conditions is then plotted in Figure 11 for the pre-learning, post-learning, and offline-learning controllers. In the figure, all system states except the pitch angle are set to 0 for the initial condition. The cost is obtained for the initial conditions with a range of pitch angle (x_2). The figure confirms that the trained controller has superior performance over the un-trained one with much smaller cost for the same initial condition, and that the proposed controller after learning provides generally identical control to the one trained offline.

6. Conclusions

It is shown in stability analysis that using the proposed MVFA to provide alternative realisation of the single-critic configuration for SPI is feasible and effective. The pro-

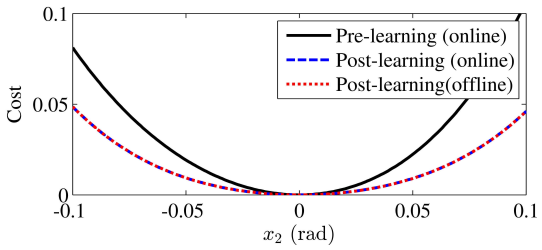

Figure 11. Cost comparison for suppressing flutter triggered by initial conditions.

posed method eliminates the need for stabilising mechanisms in either the critic or actor NN tuning, without jeopardising closed-loop stability, and without complicating the problem, as confirmed in theoretical proof and demonstrated in numerical studies. In general, the proposed MVFA used in a single-critic configuration for SPI, together with the study on parameters convergence and closed-loop stability, serve as a new development to the online SPI theory framework.

It is worth noting that the proposed MVFA scheme is model-based. Many successful model-free applications (Abouheaf, Gueaieb, & Sharaf, 2018; Luo, Wu, & Huang, 2018; Radac, Precup, & Roman, 2018) have shed a light in future works on advanced model-free MVFA based schemes that: (1) features better adaptability and robustness in circumstances with unknown, uncertain or time-varying system dynamics; (2) delivers simplified online implementation enabled by the MVFA approach. 2.5

References

Abouheaf, M., Gueaieb, W., & Sharaf, A. (2018, Oct). Model-free adaptive learning control scheme for wind turbines with doubly fed induction generators. *IET Renewable Power Generation*, *12*(14), 1675-1686.

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

Al-Tamimi, A., Lewis, F. L., & Abu-Khalaf, M. (2008). Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *38*(4), 943-949.

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

Bellman, R. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

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

Feng, T., Zhang, H., Luo, Y., & Zhang, J. (2015). Stability analysis of heuristic dynamic programming algorithm for nonlinear systems. *Neurocomputing*, *149, Part C*, 1461-1468.

Finlayson, B. A. (1972). *The method of weighted residuals and variational principles: With application in fluid mechanics, heat and mass transfer*. New York, NY: Academic Press.

Heydari, A. (2014). Revisiting approximate dynamic programming and its convergence. *IEEE Transactions on Cybernetics*, *44*(12), 2733-2743.

Heydari, A., & Balakrishnan, S. N. (2013). Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(1), 145-157.

Heydari, A., & Balakrishnan, S. N. (2014). An adaptive critic-based scheme for consensus control of nonlinear multi-agent systems. *International Journal of Control*, *87*(12), 2463-

2474.
Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366.
Huang, Y., Wang, D., & Liu, D. (2017). Bounded robust control design for uncertain nonlinear systems using single-network adaptive dynamic programming. *Neurocomputing*, *266*, 128-140.
Ioannou, P. A., & Sun, J. (1996). *Robust adaptive control*. Upper Saddle River, NJ: PTR Prentice-Hall.
Jiang, H., & He, H. (2018). Data-driven distributed output consensus control for partially observable multiagent systems. *IEEE Transactions on Cybernetics*.
Jiang, Y., & Jiang, Z.-P. (2014). Robust adaptive dynamic programming and feedback stabilization of nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 882-893.
Jiang, Y., & Jiang, Z.-P. (2015). Global adaptive dynamic programming for continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, *60*(11), 2917-2929.
Jiang, Z.-P., & Jiang, Y. (2013). Robust adaptive dynamic programming for linear and nonlinear systems: An overview. *European Journal of Control*, *19*(5), 417-425.
Kiumarsi, B., & Lewis, F. L. (2015). Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(1), 140-151.
Kiumarsi, B., Lewis, F. L., & Levine, D. S. (2015). Optimal control of nonlinear discrete time-varying systems using a new neural network approximation structure. *Neurocomputing*, *156*, 157-165.
Lewis, F. L., Jagannathan, S., & Yesildirek, A. (1999). *Neural network control of robot manipulators and nonlinear systems*. London: Taylor & Francis.
Li, J., Modares, H., Chai, T., Lewis, F. L., & Xie, L. (2017). Off-policy reinforcement learning for synchronization in multiagent graphical games. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2434-2445.
Liu, D., Huang, Y., Wang, D., & Wei, Q. (2013). Neural-network-observer-based optimal control for unknown nonlinear systems using adaptive dynamic programming. *International Journal of Control*, *86*(9), 1554-1566.
Liu, D., Wang, D., Wang, F.-Y., Li, H., & Yang, X. (2014). Neural-network-based online HJB solution for optimal robust guaranteed cost control of continuous-time uncertain nonlinear systems. *IEEE Transactions on Cybernetics*, *44*(12), 2834-2847.
Liu, D., Yang, X., & Li, H. (2013). Adaptive optimal control for a class of continuous-time affine nonlinear systems with unknown internal dynamics. *Neural Computing and Applications*, *23*(7-8), 1843-1850.
Lopez, V. G., Sanchez, E. N., Alanis, A. Y., & Rios, J. D. (2017). Real-time neural inverse optimal control for a linear induction motor. *International Journal of Control*, *90*(4), 800-812.
Luo, B., Liu, D., Huang, T., & Wang, D. (2016). Model-free optimal tracking control via critic-only Q-learning. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(10), 2134-2144.
Luo, B., Wu, H.-N., & Huang, T. (2018, June). Optimal output regulation for model-free Quanser helicopter with multistep Q-Learning. *IEEE Transactions on Industrial Electronics*, *65*(6), 4953-4961.
Luo, B., Wu, H.-N., Huang, T., & Liu, D. (2015). Reinforcement learning solution for HJB equation arising in constrained optimal control problem. *Neural Networks*, *71*, 150-158.
Luy, N. T. (2018). Distributed cooperative H_∞ optimal tracking control of MIMO nonlinear multi-agent systems in strict-feedback form via adaptive dynamic programming. *International Journal of Control*, *91*(4), 952-968.
Lv, Y., Na, J., & Ren, X. (2017). Online H_∞ control for completely unknown nonlinear systems via an identifier–critic-based ADP structure. *International Journal of Control*.
Lv, Y., Na, J., Yang, Q., Wu, X., & Guo, Y. (2016). Online adaptive optimal control for

continuous-time nonlinear systems with completely unknown dynamics. *International Journal of Control*, *89*(1), 99-112.

Modares, H., & Lewis, F. L. (2014). Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. *Automatica*, *50*(7), 1780-1792.

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

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

Modares, H., Naghibi Sistani, M.-B., & Lewis, F. L. (2013). A policy iteration approach to online optimal control of continuous-time constrained-input systems. *ISA Transactions*, *52*(5), 611-621.

Mu, C., Wang, D., & He, H. (2017). Novel iterative neural dynamic programming for data-based approximate optimal control design. *Automatica*, *81*, 240-252.

Mu, C., Wang, D., & He, H. (2018). Data-driven finite-horizon approximate optimal control for discrete-time nonlinear systems using iterative HDP approach. *IEEE Transactions on Cybernetics*, *48*(10), 2948-2961.

Na, J., & Herrmann, G. (2014). Online adaptive approximate optimal tracking control with simplified dual approximation structure for continuous-time unknown nonlinear systems. *IEEE/CAA Journal of Automatica Sinica*, *1*(4), 412-422.

Radac, M.-B., Precup, R.-E., & Roman, R.-C. (2018, Feb). Data-driven model reference control of MIMO vertical tank systems with model-free VRFT and Q-Learning. *ISA Transactions*, *73*, 227-238.

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

Song, R., Lewis, F. L., Wei, Q., & Zhang, H. (2016). Off-policy actor-critic structure for optimal control of unknown systems with disturbances. *IEEE Transactions on Cybernetics*, *46*(5), 1041-1050.

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

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

Vamvoudakis, K. G., & Lewis, F. L. (2012). Online solution of nonlinear two-player zero-sum games using synchronous policy iteration. *International Journal of Robust and Nonlinear Control*, *22*(13), 1460-1483.

Vamvoudakis, K. G., Vrabie, D., & Lewis, F. L. (2014). Online adaptive algorithm for optimal control with integral reinforcement learning. *International Journal of Robust and Nonlinear Control*, *24*(17), 2686-2710.

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

Škach, J., Kiumarsi, B., Lewis, F. L., & Straka, O. (2018). Actor-critic off-policy learning for optimal control of multiple-model discrete-time systems. *IEEE Transactions on Cybernetics*, *48*(1), 29-40.

Wang, D., He, H., & Liu, D. (2017a). Adaptive critic nonlinear robust control: A survey. *IEEE Transactions on Cybernetics*, *47*(10), 3429-3451.

Wang, D., He, H., & Liu, D. (2017b). Improving the critic learning for event-based nonlinear H_∞ control design. *IEEE Transactions on Cybernetics*, *47*(10), 3417-3428.

Wang, D., & Liu, D. (2013). Neuro-optimal control for a class of unknown nonlinear dynamic systems using SN-DHP technique. *Neurocomputing*, *121*, 218-225.

Wang, D., Liu, D., Li, H., & Ma, H. (2014). Neural-network-based robust optimal control design for a class of uncertain nonlinear systems via adaptive dynamic programming. *Information Sciences*, *282*, 167-179.

Wang, D., Liu, D., Wei, Q., Zhao, D., & Jin, N. (2012). Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming. *Automatica*, *48*(8), 1825-1832.

Wang, D., Liu, D., Zhang, Q., & Zhao, D. (2016). Data-based adaptive critic designs for nonlinear robust optimal control with uncertain dynamics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *46*(11), 1544-1555.

Wang, D., Mu, C., Yang, X., & Liu, D. (2017). Event-based constrained robust control of affine systems incorporating an adaptive critic mechanism. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *47*(7), 1602-1612.

Wang, F.-Y., Zhang, H., & Liu, D. (2009). Adaptive dynamic programming: An introduction. *IEEE Computational Intelligence Magazine*, *4*(2), 39-47.

Wang, Z., Behal, A., & Marzocca, P. (2011). Model-free control design for multi-input multi-output aeroelastic system subject to external disturbance. *Journal of Guidance, Control, and Dynamics*, *34*(2), 446-458.

Wei, Q., Lewis, F. L., Sun, Q., Yan, P., & Song, R. (2017). Discrete-time deterministic Q-learning: A novel convergence analysis. *IEEE Transactions on Cybernetics*, *47*(5), 1224-1237.

Wei, Q., & Liu, D. (2014). A novel iterative θ-adaptive dynamic programming for discrete-time nonlinear systems. *IEEE Transactions on Automation Science and Engineering*, *11*(4), 1176-1190.

Wei, Q., Liu, D., & Lin, H. (2016). Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems. *IEEE Transactions on Cybernetics*, *46*(3), 840-853.

Wei, Q., Wang, F.-Y., Liu, D., & Yang, X. (2014). Finite-approximation-error-based discrete-time iterative adaptive dynamic programming. *IEEE Transactions on Cybernetics*, *44*(12), 2820-2833.

Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences* (Ph.D. Thesis). Harvard University.

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

Zhang, H., Cui, L., & Luo, Y. (2013). Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network ADP. *IEEE Transactions on Cybernetics*, *43*(1), 206-216.

Zhao, D., Xia, Z., & Wang, D. (2015). Model-free optimal control for affine nonlinear systems with convergence analysis. *IEEE Transactions on Automation Science and Engineering*, *12*(4), 1461-1468.

Zhong, X., He, H., Wang, D., & Ni, Z. (2018). Model-free adaptive control for unknown nonlinear zero-sum differential game. *IEEE Transactions on Cybernetics*, *48*(5), 1633-1646.