Asymptotically-Stable Adaptive-Optimal Control Algorithm with Saturating Actuators and Relaxed Persistence of Excitation

Kyriakos G. Vamvoudakis[1], Member, IEEE   Marcio F. Miranda[2]   João P. Hespanha[1], Fellow, IEEE

Abstract— **This paper proposes a control algorithm based on adaptive dynamic programming to solve the infinite-horizon optimal control problem for known deterministic nonlinear systems with saturating actuators and non-quadratic cost functionals. The algorithm is based on an actor/critic framework where a critic neural network is used to learn the optimal cost and an actor neural network is used to learn the optimal control policy. The adaptive control nature of the algorithm requires a persistence of excitation condition to be a priori validated, but this can be relaxed by using previously stored data concurrently with current data in the update of the critic neural network. A robustifying control term is added to the controller to eliminate the effect of residual errors, leading to asymptotically stability of the closed-loop system. Simulation results show the effectiveness of the proposed approach for a controlled Van-der Pol oscillator and also for a power systems plant.**

Index Terms— **Saturating actuators, approximate dynamic programming, asymptotic stability, optimal control, reinforcement learning.**

I. Introduction

Optimal control deals with the problem of finding a control law for a given system and user defined optimality criterion. It can be derived using either Pontryagin's maximum principle (a necessary condition), or by solving the Hamilton-Jacobi-Bellman (HJB) equation (a sufficient condition). However, either approach is typically intractable. Adaptive control techniques on the other side are designed for online use but cannot typically optimize user-defined performance indices.

Adaptive dynamic programming techniques were proposed by Werbos [1, 2] and bring together the advantages of adaptive and optimal control to obtain approximate and forward in time solutions to difficult optimization problems [3–5]. But all the existing algorithms — such as the ones developed in [6–10] and most of the references therein — can only guarantee uniform ultimate boundedness of the closed-loop system, i.e. a "milder" form of stability [11] in the sense of Lyapunov, and require an a-priori knowledge of a persistence of excitation condition.

The need for adaptive controllers with the ability to learn optimal solutions, while still guaranteeing asymptotic stability motivates our research. The algorithm proposed uses the structure of a reinforcement learning algorithm called Policy Iteration (PI), which is inspired by behavioral psychology [12]. This two step algorithm has an actor/critic structure which involves two neural networks (NNs): one, the critic NN, is trained to become an approximation of the Value function solution at the policy evaluation step, while the second one is trained to approximate an optimal policy at the policy improving step.

In industrial applications, physical inputs to devices (such as voltages, currents, flows, and torques) [12] are subject to saturations, which must be considered in the optimal control problem. To the best of our knowledge, there are no *asymptotically stable online solutions* to the continuous time HJB equation with saturations since they add addition nonlinearities to the HJB and make the problem more difficult. This challenge is also addressed here.

Related work

In [13], the authors propose a nonquadratic functional that involves bounded control inputs but do not provide solutions to the HJB equation and based on that, the authors in [14] used neural networks to solve the HJB offline. A novel iterative two-stage dual heuristic programming (DHP) to solve the optimal control problem for a class of switched discrete-time systems subject to actuators saturation has been introduced in [15].

The work of [16] develops a learning framework for computing the HJB solution of discrete time systems under bounded disturbances, but the authors only prove uniform ultimate boundedness. In [17] the authors derive a gain condition for global asymptotic stability that allows the presence of dynamic uncertainties with unmeasured state and unknown system order/dynamics. However, the policy iteration algorithm is not performed in a synchronous manner, i.e. here the policy evaluation and policy improvement steps take place simultaneously. Recently the authors in [18] proposed an approximate dynamic programming method to solve the optimal control problem for complex-valued systems. Several adaptive/critic designs have been applied in real applications, such as [19], where the authors propose an intelligent adaptive/critic controller to provide a nonlinear optimal control to a smart grid operation in an environment with high short-term uncertainty and variability.

Temporal Difference (TD) algorithms [12], [20], incrementally update the policies based on each individual ex-

[1] K. G. Vamvoudakis, and J. P. Hespanha are with the Center for Control, Dynamical-systems and Computation (CCDC), University of California, Santa Barbara, CA 93106-9560 USA e-mail: kyriakos@ece.ucsb.edu, hespanha@ece.ucsb.edu.

[2] M. F. Miranda is with the COLTEC, Universidade Federal de Minas Gerais, Belo Horizonte - MG CEP 31270-901, Brazil e-mail: marcio.fantini@gmail.com.

*This material is based upon work supported in part by ARO MURI Grant number W911NF0910553, ARO grant W911NF-09-D-0001 (Inst. for Collaborative Biotechnologies) and by grant CAPES No BEX-9586/11-3.

perience. The combination of experience replay [21, 22] and fitted Q iteration [23, 24], with on-line TD methods have been shown in the past to speed up learning in low-dimensional control problems. But again, due to the large state space, value function approximation is a necessity, violating the assumptions for guaranteed convergence and thus leaving room for asymptotic performance gains as well. The authors in [25, 26] have shown that DHP algorithms can eventually find an optimal solution without the explicit need for stochastic exploration, but the value learning algorithms (i.e. TD, TD(0)) could not. Especially the work of [26] states that for any given plant, data from optimal and suboptimal action strategies can be used without difficulty in system identification which means that DHP has less excitation problems. There is an extensive research work on Markov Decision Processes (MDPs) and *linear discrete-time systems* that is not readily applied to control systems, does not take into account input constraints and does not have proper convergence proofs. In this direction, the effort of "residual gradients" [27] or "Galerkinized" methods [26, 28, 29], where the critic NN weights do not converge to the optimal values in the stochastic case [30, 31], but converge only in the deterministic case. The aforementioned limitations of "Galerkinized" methods have been overcome in [32] where the authors solved this problem under standard technical assumptions. In [33] the authors propose a variant of DHP that has a guaranteed convergence, under certain smoothness conditions and a greedy policy, when using a general smooth nonlinear function approximator for the critic.

In MDPs, it is known that conventional TD methods do not use trajectory data efficiently, since after a gradient update, the state transitions and rewards are ignored. The Least Squares Temporal Difference Learning algorithm (LSTD) [34] on the other side, is known for its efficient use of sample experiences compared to pure TD. In spite of the fact that it is initially appealing to attempt to use LSTD in the evaluation step of a policy iteration algorithm, this combination can be problematic. LSTD is more accurate than TD(0), but more algorithmically complex. This can be easily seen when the value function approximator involves a very large number of basis sets. The LSTD algorithm is very similar to Heuristic Dynamic Programming (HDP) [35], but it involves extra advantages that include, convergence with arbitrary initial conditions and recursive formulation.

The authors in [36] propose a generalized value iteration algorithm for discrete-time systems to overcome the disadvantage of traditional value iteration algorithms by allowing an arbitrary positive semi-definite function as initialization. Our work in this paper is focused on a "Galerkinized" asymptotically stable algorithm for known deterministic nonlinear *continuous-time systems*.

There is a lot of research work on approximate dynamic programming for linear systems [26, 37] where the persistence of excitation condition used also in adaptive control [38] needs to be satisfied in order to guarantee convergence of the critic NN. For nonlinear systems, there is no clear development. Recent adaptive optimal control algorithms with approximate dynamic programming [6, 8, 39] require a persistence of excitation (PE) condition that is essentially analogous to space exploration in reinforcement learning [12]. This condition is restrictive and often difficult to guarantee in practice. Hence, convergence cannot be guaranteed. The work of [40] from the adaptive control side, and the works of [41] and [42] from the reinforcement learning side propose some frameworks that rely on concurrently using *current and recorded data* for adaptation to obviate the difficulty of guaranteeing convergence with PE. Recently the authors in [43] have used concurrent learning in optimal adaptive control but they *only prove a "milder" form of stability, namely uniform ultimate boundedness of the closed-loop signals* by using an approach that is based on integral reinforcement learning.

Contributions

The contributions of this paper lie in the development of an adaptive learning algorithm to solve an infinite horizon optimal control problem for known deterministic nonlinear systems, while taking into account symmetric input constraints. The algorithm proposed is an appropriate combination of adaptive control, optimization and reinforcement learning. A novelty of our approach lies in the use of *concurrent* information to relax the persistence of excitation condition. By "concurrent" it is meant that current and stored data is used for the adaptation process, which facilitates the convergence of the algorithm. In fact, we prove asymptotic stability of the closed-loop system, which includes the online updates of the critic and the actor neural networks using state measurements.

The paper is structured as follows. In Section II we formulate the optimal control problem with saturated inputs. The approximate solution for the HJB equation is presented in Section III. The proof of asymptotic stability of the closed loop is presented in Section IV. Simulation results for a power system plant and a controlled Van-der Pol oscillator are given in Section V. Finally, Section VI concludes and discusses about future work.

Notation

The notation used here is standard: $\mathbb{R}^+$ is the set of positive real numbers and $\mathbb{Z}^+$ is the set of positive integer numbers. The superscript $\star$ is used to denote the optimal solution of an optimization, $\lambda_{\min}(A)$ is the minimum eigenvalue of a matrix A and $\mathbf{1}_m$ is the column vector with m ones. The gradient of a scalar-valued function with respect to a vector-valued variable x is defined as a column vector, and is denoted by $\nabla \coloneqq \partial/\partial x$. A function $\alpha : \mathbb{R}^+ \to \mathbb{R}$ is said to belong to class $\mathcal{K}(\alpha \in \mathcal{K})$ functions if it is continuous, strictly increasing and $\alpha(0) = 0$.

II. Problem formulation

Consider the nonlinear continuous-time system given by

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t); \quad x(0) \coloneqq x_0, \ t \geqslant 0 \qquad (1)$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in U \subseteq \mathbb{R}^m$ is the control input, $f(x(t)) \in \mathbb{R}^n$ and $g(x(t)) \in \mathbb{R}^{n\times m}$ are known functions. We assume that the $x(t)$ is available for full state feedback.

It is desired to minimize the following infinite horizon cost functional

$$V(x(0)) = \int_0^{\infty} r\left(x(\tau), u(\tau)\right) d\tau, \ \forall x(0) \tag{2}$$

with

$$r(x,u) \quad = \quad Q(x) + R_s(u), \ \forall x, u \tag{3}$$

with functions $Q(x)$ positive definite and $R_s(u)$ non-negative $\forall u \in U$. Since the present paper is concerned with providing an asymptotically stable framework for solving the HJB equation we will provide the following definition.

Definition 1: [11] Given an autonomous, time-invariant nonlinear system of the form $\dot{x}(t) = f(x(t))$ with $x(t) \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^n$ and f continuous and an equilibrium point given as x_e, i.e. $f(x_e) = 0$. Let $\psi(t; 0, \bar{x})$ denote the unique solution $x(t)$ to $\dot{x}(t) = f(x(t))$ that corresponds to $x(0) = \bar{x}$. Then the equilibrium point x_e is said to be asymptotically stable, if $\forall \epsilon \in \mathbb{R}^+$ there exists a $\delta \in \mathbb{R}^+$ such that,

$$\bar{x} \in \mathcal{B}(x_e, \delta) \Rightarrow \begin{cases} \psi(t; 0, \bar{x}) \in \mathcal{B}(x_e, \epsilon), \ \forall t \geqslant 0 \\ \lim_{t\to\infty} \psi(t; 0, \bar{x}) = x_e, \end{cases}$$

where $\mathcal{B}(\bar{x}, \epsilon)$ denotes the open ball centered at $\bar{x}$ of radius ϵ, i.e. the set, $\{x \in \mathbb{R}^n : \|x - \bar{x}\| < \epsilon\}$. □

The optimal control problem is to find an admissible control $u^\star(t)$ such that the equilibrium point of the closed-loop system (1) is asymptotically stable on $\mathbb{R}^n$ in the sense of Definition 1 and the value V is finite. To force *bounded inputs*, (e.g. $|u_i| \leqslant \bar{u}, \forall i \in \{1, \ldots, m\}$) we follow the approach in [13] and use a nonquadratic penalty function of the form,

$$R_s(u) = 2\sum_{i=1}^{m} \int_0^{u_i} \left(\theta^{-1}(v_i)\right)^T \varrho_i dv_i, \forall u,$$

with weighting factor $\varrho_i \in \mathbb{R}^+$, $i = \{1, \ldots, m\}$ and with abuse of notation we can write the component-wise operations in compact form as

$$R_s(u) = 2\int_0^{u} \left(\theta^{-1}(v)\right)^T R dv, \ \forall u,$$

where R is a diagonal positive definite matrix consisting of the $\varrho_i > 0$, $i = \{1, \ldots, m\}$ terms, $v \in \mathbb{R}^m$, and $\theta(\cdot)$ is a continuous, one-to-one real-analytic integrable function of class C^μ, $\mu \geqslant 1$, used to map $\mathbb{R}$ onto the interval $(-\bar{u}, \bar{u})$ satisfying $\theta(0) = 0$. Also note that $R_s(u)$ is positive definite because $\theta^{-1}(v)$ is monotonic odd. E.g., one could select

$$\begin{aligned} R_s(u) &= 2\int_0^{u} \left(\theta^{-1}(v)\right)^T R dv \\ &\coloneqq 2\int_0^{u} \left(\bar{u}\tanh^{-1}(v/\bar{u})\right)^T R dv > 0, \forall u. \end{aligned} \tag{4}$$

The optimal value function is defined as,

$$V^\star(x(t)) = \min_{u\in U} \int_t^{\infty} r(x, u) d\tau, \ \forall x, \ t \geqslant 0, \tag{5}$$

subject to the state dynamics in (1). The Hamiltonian of (1) associated with the cost function (2)-(3), can be written as,

$$\begin{aligned} H(x, u, \nabla V(x)) = \nabla V(x)^T \big(f(x) + g(x)u\big) \\ + Q(x) + R_s(u), \ \forall x, u. \end{aligned} \tag{6}$$

The constrained optimal control $\bar{\mathcal{K}}^*(x)$ (i.e. $u(t) := \bar{\mathcal{K}}^*(x)$) for the system (1), with cost (3), (4), (5), can be obtained using the stationarity condition in the Hamiltonian (6):

$$\begin{aligned} \bar{\mathcal{K}}^*(x) &= \arg\min_u H(x, u, \nabla V(x)) \\ &\Rightarrow \nabla V^\star(x)^T g(x) + 2R\theta^{-T}\big(\bar{\mathcal{K}}^*(x)\big) = 0 \\ &\Rightarrow \bar{\mathcal{K}}^\star(x) = -\theta\left(\frac{1}{2}R^{-1}g^T(x)\nabla V^\star(x)^T\right), \ \forall x. \end{aligned} \tag{7}$$

The corresponding optimal cost and optimal control satisfy the following HJB equation,

$$\begin{aligned} H^\star(x, \bar{\mathcal{K}}^\star(x), \nabla V^\star(x)) \coloneqq \nabla V^\star(x)^T \big(f(x) + g(x)\bar{\mathcal{K}}^\star(x)\big) \\ + Q(x) + R_s(\bar{\mathcal{K}}^\star(x)) = 0, \ \forall x. \end{aligned} \tag{8}$$

The next result provides a sufficient condition for the existence of the optimal control solution.

Theorem 1: Suppose there exists a positive definite and radially unbounded smooth function $V \in C^1$ that satisfies $V(0) = 0$ and

$$H(x, \bar{\mathcal{K}}^\star(x), \nabla V(x)) = 0, \ \forall x \tag{9}$$

with $H(\cdot)$ given by (8) and $\bar{\mathcal{K}}^\star(x)$ given by

$$\bar{\mathcal{K}}^\star(x) = -\theta\left(\frac{1}{2}R^{-1}g^T(x)\nabla V(x)^T\right), \tag{10}$$

and that the closed-loop system (1) and (10) has a locally Lipschitz right hand size with state $x \in \mathbb{R}^n$, i.e. $x \mapsto f(x) + g(x)\bar{\mathcal{K}}^\star(x)$. The origin is a globally asymptotically stable equilibrium point of the closed-loop system (1) with control (10) and the control policy (10) minimizes the cost (5).

Proof of Theorem 1. Because of (6) and (9) (since we have assumed that V is a positive definite and radially unbounded function that solves the HJB), the time derivative of V along closed-loop solutions satisfies

$$\begin{aligned} \dot{V} = \nabla V(x)^T(f(x) + g(x)\bar{\mathcal{K}}^\star(x)) = -R_s(\bar{\mathcal{K}}^\star(x)) - Q(x) \\ \leqslant -Q(x), \end{aligned}$$

where we used (9) with $\bar{\mathcal{K}}^\star(x)$ given by (10). Using V as a Lyapunov function, we conclude that the origin is a globally asymptotically stable equilibrium point of (1).

Now since V is smooth and $V(0) = 0$, as $t \to \infty$, then one has,

$$V(x(0)) + \int_0^{\infty} \nabla V(x)^T \left(f(x) + g(x)\bar{\mathcal{K}}^\star(x)\right) = 0. \tag{11}$$

Since the function $V(x)$ is smooth, converge to zero as $t \to \infty$ (due to asymptotic stability) and $V(0) = 0$, using (11) we can write (2) as,

$$V(x(0);u) = \int_0^\infty \big(R_s(u) + Q(x)\big)dt + V^\star(x(0)) \\ + \int_0^\infty \nabla V^\star(x)^T(f(x) + g(x)\bar{\mathcal{K}}^\star(x))dt, \ \forall u.$$

By subtracting zero (using the HJB equation) we have,

$$V(x(0);u) = \int_0^\infty \bigg(\big(R_s(u) - R_s(\bar{\mathcal{K}}^\star(x))\big) \\ + \nabla V^\star(x)^T g(x)(u - \bar{\mathcal{K}}^\star(x))\bigg)dt + V^\star(x(0)), \ \forall u. \quad (12)$$

By noting that, $\nabla V^\star(x)^T g(x) = -2R\theta^{-T}(\bar{\mathcal{K}}^\star(x))$ in (12) and after completing the squares we have,

$$V(x(0);u) = \int_0^\infty \bigg(\big(2(\int_0^u (\theta^{-1}(v))^T dv - \int_0^{\bar{\mathcal{K}}^\star(x)} (\theta^{-1}(v))^T dv)\big) \\ - R\theta^{-T}(u)(u - \bar{\mathcal{K}}^\star(x))\bigg)dt + V^\star(x(0)), \ \forall u.$$

We can complete the squares and hence we have,

$$V(x(0);u) = \int_0^\infty R_s(u - \bar{\mathcal{K}}^\star(x))dt + V^\star(x(0)), \ \forall u.$$

Now by setting $u := \bar{\mathcal{K}}^\star(x)$ one can show that,

$$V^\star(x(0)) \leqslant V(x(0);u),$$

from which the result follows. ■

The following section provides approximate techniques to converge to the solution of the HJB equation (8).

III. Approximate Solution

The structure used for our approximate solution is motivated by the Policy Iteration Algorithm that follows, where ϵ_{ac} is a small number used to terminate the algorithm when two consecutive value functions differ by less than ϵ_{ac}. In the linear case, this algorithm reduces to Kleinman's algorithm [44].

Algorithm 1: Policy Iteration for Nonlinear Systems

```
1: procedure
2:     Given admissible policies $\mu^{(0)}$ and $i = 1$
3:     while $\|V^{\mu^{(i)}} - V^{\mu^{(i-1)}}\| \geqslant \epsilon_{\mathrm{ac}}, \forall x$ do
4:         Solve for the value $V^{(i)}(x)$ using Bellman's equation
               $Q(x)+\nabla V^{\mu^{(i)}T}(f(x)+g(x)\mu^{(i)})+R_s(\mu^{(i)}) = 0, \ V^{\mu^{(i)}}(0) = 0,$
5:         Update the control policy $\mu^{(i+1)}$ using
               $\mu^{(i+1)} = -\theta\Big(\frac{1}{2}R^{-1}g^T(x)\nabla V^{\mu^{(i)}T}\Big)$
6:         $i := i + 1$
7:     end while
8: end procedure
```

The next subsection lays the foundation for updating the two steps 4 and 5 in Policy Iteration *simultaneously* by using data collected along the closed-loop trajectory.

A. Critic neural network and recorded past data

The first step to solve the HJB equation (8) is locally to approximate the value function $V^\star(x)$ in (5) with a critic neural network (NN), within a set $\Omega \subseteq \mathbb{R}^n$ that contains the origin, as follows

$$V^\star(x) = W^{\star T}\phi(x) + \epsilon(x), \ \forall x \quad (13)$$

where $W^\star \in \mathbb{R}^N$ is an ideal weight vector satisfying $\|W^\star\| \leqslant W_{\mathrm{m}}$; $\phi(x) : \Omega \to \mathbb{R}^N$, $\phi(x) = [\varphi_1(x) \ \varphi_2(x) \ \dots \ \varphi_N(x)]^T$ are the NN activation functions such that $\varphi_i(0) = 0$ and $\nabla\varphi_i(0) = 0$, $\forall i = 1, \dots, N$; N is the number of neurons in the hidden layer; and $\epsilon(x)$ is the NN approximation error.

One should pick the NN activation functions $\varphi_i(x)$, $\forall i \in \{1, 2, \dots, N\}$ as quadratic, radial basis or sigmoid functions so that they define a complete independent basis set for $V^\star$. In this case, $V^\star$ and its derivatives

$$\nabla V^\star(x) = \left[\frac{\partial}{\partial x}\phi(x)\right]^T W^\star + \frac{\partial}{\partial x}\epsilon(x) \\ =: \nabla\phi(x)^T W^\star + \nabla\epsilon(x), \ \forall x \in \Omega. \quad (14)$$

can be uniformly approximated on any given compact set Ω. According to Weierstrass Higher Order Approximation Theorem [45], [14], as the number of basis sets N increases, the approximation error on a compact set Ω goes to zero, i.e., $\epsilon(x) \to 0$ as $N \to \infty$. We shall require a form of uniformity in this approximation result that is common in neuro-adaptive control and other approximation techniques [38, 45]. This assumption also involves the *approximate HJB* defined by

$$H^\star(x, \bar{\mathcal{K}}^\star(x), W^{\star T}\nabla\phi) := W^\star\nabla\phi(f(x) + g(x)\bar{\mathcal{K}}^\star(x)) \\ + Q(x) + R_s(\bar{\mathcal{K}}^\star(x)) = \epsilon_H, \ \forall x, \quad (15)$$

which is obtained by using (14) in (8) and that leads to the residual error

$$\epsilon_H := H^\star(\cdot) - H(\cdot) = -\nabla\epsilon^T(f(x) + g(x)\bar{\mathcal{K}}^\star(x)), \ \forall x, \quad (16)$$

where for brevity we have omitted the arguments of H and $H^\star$.

Assumption 1 (Critic Uniform Approximation): The critic activation functions ϕ, the value function approximation error ϵ, their derivatives, and the Hamiltonian residual error ϵ_H are all uniformly bounded on a set $\Omega \subseteq \mathbb{R}^n$, in the sense that there exist finite constants $\phi_{\mathrm{m}}, \phi_{\mathrm{dm}}, \epsilon_{\mathrm{m}}, \epsilon_{\mathrm{dm}}, \epsilon_{\mathrm{Hm}} \in \mathbb{R}^+$ such that $|\phi(x)| \leqslant \phi_{\mathrm{m}}$, $|\nabla\phi(x)| \leqslant \phi_{\mathrm{dm}}$, $|\epsilon(x)| \leqslant \epsilon_{\mathrm{m}}$, $|\nabla\epsilon(x)| \leqslant \epsilon_{\mathrm{dm}}$, $|\epsilon_H(x)| \leqslant \epsilon_{\mathrm{Hm}}$, $\forall x \in \Omega$. In order to get ϵ small we also assume that we have a large number of basis sets, i.e. $N \to \infty$. □

Since the ideal weights $W^\star$ for the (approximate) value function $V^\star(x)$ that appear in (13) are unknown, one must consider the *critic weight estimates* $\hat{W} \in \mathbb{R}^N$, associated with the approximate value function:

$$\hat{V}(x) = \hat{W}^T\phi(x), \forall x. \quad (17)$$

Our objective is to find an update law for the weight estimates $\hat{W}$ so that they converge to the ideal values $W^\star$, and thus provide a good estimate

$$\hat{H}(x, u, \hat{W}^T \nabla\phi) := \hat{W}^T \nabla\phi(f(x) + g(x)u) + Q(x) + R_s(u), \ \forall x, u, \quad (18)$$

for the (approximate) Hamiltonian.

Definition 2: [38] A vector signal $\Phi(t)$ is *exciting* over the interval $[t, t + T_{\mathrm{PE}}]$, with $T_{\mathrm{PE}} \in \mathbb{R}^+$ if there exists $\beta_1, \beta_2 \in \mathbb{R}^+$ such that $\beta_1 I \leqslant \int_t^{t+T} \Phi(\tau)\Phi^T(\tau)d\tau \leqslant \beta_2 I, \forall t$ with I an identity matrix of appropriate dimensions. □

There is a need to develop a learning framework to find a tuning law for $\hat{W}$ in order to achieve convergence of (18) to the (approximate) Hamiltonian (15) along the closed-loop trajectories. But in order to attain that, one would typically need persistency of excitation (see Definition 2) for the vector $\omega(t)$ defined by

$$\omega(t) := \nabla\phi\big(x(t)\big)\Big(f\big(x(t)\big) + g\big(x(t)\big)u(t)\Big), \quad (19)$$

along the closed-loop trajectories [38]. To weaken the need to guarantee a-priori, a persistency of excitation condition in the sense of Definition 2 for infinite-time, we follow the approach proposed in [46] that uses *past recorded data, concurrently with current data*. To this effect, we define the Hamiltonian error corresponding to the data collected at the *current time* t:

$$\begin{aligned} e(t) &:= \hat{H}\Big(x(t), u(t), \hat{W}(t)^T \nabla\phi\big(x(t)\big)\Big) - H^\star(x, \bar{\mathcal{K}}^\star(x), \nabla V^\star) \\ &= \hat{H}\Big(x(t), u(t), \hat{W}(t)^T \nabla\phi\big(x(t)\big)\Big), \forall x, u \end{aligned}$$

where the latter equality is due to (8), and the error corresponding to data *previously collected at times* $t_0, t_1, \ldots, t_k < t$,

$$\begin{aligned} e_{\mathrm{buff}_i}(t_i, t) &:= \hat{H}\Big(x(t_i), u(t_i), \hat{W}(t)^T \nabla\phi\big(x(t_i)\big)\Big) \\ &:= \hat{W}(t)^T \nabla\phi(x(t_i)(f(x(t_i)) + g(x(t_i))u(t_i)) \\ &\quad + Q(x(t_i)) + R_s(u(t_i)). \end{aligned}$$

We draw attention to the reader that, while the error $e_{\mathrm{buff}_i}(t_i, t)$ uses past *state and input data* $x(t_i)$ and $u(t_i)$, respectively, it is defined based on the *current* weight estimates $\hat{W}(t)$.

The current and previous errors defined above can be combined into the following (normalized) global error

$$E(t) = \frac{1}{2}\bigg(\frac{e(t)^2}{(\omega(t)^T\omega(t) + 1)^2} + \sum_{i=1}^{k} \frac{e^2_{\mathrm{buff}_i}(t_i, t)}{(\omega(t_i)^T\omega(t_i) + 1)^2}\bigg), \ \forall t,$$

where $\omega(t_i) := \nabla\phi\big(x(t_i)\big)\Big(f\big(x(t_i)\big) + g\big(x(t_i)\big)u(t_i)\Big)$.

The tuning for the critic NN is obtained by a gradient-descent-like rule as follows:

$$\begin{aligned} \dot{\hat{W}} &= -\alpha\frac{\partial E}{\partial \hat{W}} \\ &= -\alpha\frac{\omega(t)e(t)}{(\omega(t)^T\omega(t) + 1)^2} - \alpha\sum_{i=1}^{k}\frac{\omega(t_i)e_{\mathrm{buff}_i}(t_i, t)}{(\omega(t_i)^T\omega(t_i) + 1)^2} \\ &= -\alpha\frac{\omega(t)\big(\omega(t)^T\hat{W}(t) + R_s(u(t)) + Q(x(t))\big)}{(\omega(t)^T\omega(t) + 1)^2} \\ &\quad - \alpha\sum_{i=1}^{k}\frac{\omega(t_i)\big(\omega(t_i)^T\hat{W}(t) + Q(x(t_i)) + R_s(u(t_i))\big)}{(\omega(t_i)^T\omega(t_i) + 1)^2}, \end{aligned} \quad (20)$$

$\forall t > t_i \geqslant 0$, where $\alpha > 0$ is a constant gain that determines the speed of convergence. Defining the weight estimation error of the critic by

$$\tilde{W} := W^\star - \hat{W} \in \mathbb{R}^N. \quad (21)$$

We conclude from (20) that the error dynamics can be written as,

$$\dot{\tilde{W}} = -N_{\mathrm{om}} + P_{ert} \quad (22)$$

where,

$$\begin{aligned} N_{\mathrm{om}}(t) := \alpha\bigg(&\frac{\omega(t)\omega(t)^T}{(\omega(t)^T\omega(t) + 1)^2} \\ &+ \sum_{i=1}^{k}\frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i) + 1)^2}\bigg)\tilde{W}(t), \end{aligned} \quad (23)$$

can be viewed as defining the *nominal error dynamics* and

$$\begin{aligned} P_{ert}(t) := \alpha\bigg(&\frac{\omega(t)}{(\omega(t)^T\omega(t) + 1)^2}\epsilon_H(t) \\ &+ \sum_{i=1}^{k}\frac{\omega(t_i)}{(\omega(t_i)^T\omega(t_i) + 1)^2}\epsilon_H(t_i)\bigg) \end{aligned} \quad (24)$$

a *perturbation term*, bounded as $\|P_{ert}\| \leqslant \frac{\alpha}{2}(k+1)\epsilon_{\mathrm{Hm}}$, that would be zero if the Hamiltonian errors ϵ_H were absent. To derive this expression for $\dot{\tilde{W}} = -\dot{\hat{W}}$, we used (20) together with the fact that $Q(x(t)) + R_s(u(t)) = -W^{\star T}\omega(t) + \epsilon_H(t)$ which is a consequence of (8) and (16).

Theorem 2: Suppose that $\{\omega(t_1), \ldots, \omega(t_k)\}$ contains N *linearly independent vectors* and that the tuning law is given by (20). Then for every given control signal $u(t)$ we have that,

$$\frac{d\|\tilde{W}(t)\|^2}{dt} = -2\tilde{W}(t)^T N_{\mathrm{om}} \leqslant -2\alpha\lambda_{\min}(\Lambda)\|\tilde{W}\|^2 \quad (25)$$

with $\Lambda := \sum_{i=1}^{k}\frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i)+1)^2} > 0$ along solutions to (22). □

Remark 1: Typical adaptive optimal control algorithms [9] do not have the extra past-data terms $\sum_{i=1}^{k}\frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i)+1)^2}$ in the error dynamics (22) and thus need a persistence of excitation condition on $\frac{\omega(t)}{(\omega(t)^T\omega(t)+1)}$ (typically of the form $\beta_1 I \leqslant \int_t^{t+T}\frac{\omega(t)\omega(t)^T}{(\omega(t)^T\omega(t)+1)^2} \leqslant \beta_2 I$ with constants $\beta_1, \beta_2, T \in \mathbb{R}^+$) that holds for every t from $t = 0$ to $t = \infty$. This is equivalent to requiring that the matrix $\int_t^{t+T}\frac{\omega(t)\omega(t)^T}{(\omega(t)^T\omega(t)+1)^2} \in \mathbb{R}^{n\times n}$ be positive definite over any finite interval. This is equivalent to requiring that the signal $\omega(t)$ contains at least n spectral lines. This *condition cannot be verified during learning especially for nonlinear systems*. In Theorem 2, the "relaxed" persistence of excitation condition comes through the requirement that at least N of the vectors $\{\omega(t_1), \ldots, \omega(t_k)\}$ must be linearly independent, which is equivalent to the matrix Λ being positive definite. In practice, as one collects each additional vector $\omega(t_i)$, one adds a new term to the matrix Λ and one can stop recording points as soon as this matrix becomes full-rank (i.e. t_k time has been reached). From that point forward, one does not need to record new data and the assumption of Theorem 2 holds, *regardless of whether or not future data provides additional excitation*. The selection of the times t_i is somewhat arbitrary, but in our numerical simulations we typically select these values equally spaced in time. □

Remark 2: It is assumed that the maximum number of data points to be stored in the history stack (i.e., $t_0, t_1, \ldots, t_k < t$) is limited due to memory/bandwidth limitations. □

Proof of Theorem 2. Consider the following Lyapunov function

$$\mathcal{L} = \frac{1}{2\alpha}\tilde{W}(t)^T\tilde{W}(t). \tag{26}$$

By differentiating (26) along the error dynamics system trajectories one has,

$$\dot{\mathcal{L}} = -\tilde{W}(t)^T\Big(\frac{\omega(t)\omega(t)^T}{(\omega(t)^T\omega(t)+1)^2}+\Lambda\Big)\tilde{W}(t)+\tilde{W}(t)^T P_{ert}, \tag{27}$$

which is negative definite as long as,

$$\|\tilde{W}\| > \frac{\|P_{ert}\|}{\lambda_{\min}(\Lambda)}.$$

Equation (25) follows from this and the fact that $\frac{\omega(t)\omega(t)^T}{(\omega(t)^T\omega(t)+1)^2} > 0, \ \forall t$. Since $\{\omega(t_1), \ldots, \omega(t_k)\}$ has N linearly independent vectors, the matrix Λ is positive definite, from which the exponential stability of the nominal system follows. ■

B. Actor neural network

One could use a single set of weights with a sliding-mode controller as in [47] to approximate both $V^\star$ and its gradient $\nabla V^\star$, but instead we independently adjust two sets of weights: the critic weights introduced in (17) to approximate $V^\star$ and the actor weights introduced below to approximate $\nabla V^\star$ in the expression of the optimal control policy (7). While this carries additional computational burden, the flexibility introduced by this "over-parameterization" will enable us to establish convergence to the optimal solution and guaranteed Lyapunov-based stability, which seems difficult using only one set of weights.

The optimal control policy (7) can be approximated by an actor NN as follows,

$$\bar{\mathcal{K}}^\star(x) = W_u^{\star T}\phi_u(x) + \epsilon_u(x), \forall x \tag{28}$$

where $W_u^\star \in \mathbb{R}^{N_2 \times m}$ is an ideal weight matrix, $\phi_u(x)$ are the actor NN activation functions defined similarly to the critic NN, N_2 is the number of neurons in the hidden layer, and ϵ_u is the actor approximation error. As before, the NN activation functions must define a complete independent basis set so that $\bar{\mathcal{K}}^\star(x)$ can be uniformly approximated on Ω, as expressed by the following assumption.

Assumption 2 (Actor Uniform Approximation): The actor activation functions in ϕ_u and the actor residual error ϵ_u are all uniformly bounded on a set $\Omega \subseteq \mathbb{R}^n$, in the sense that there exist finite constants $\phi_{\text{um}}, \epsilon_{\text{um}} \in \mathbb{R}^+$ such that $|\phi_u(x)| \leqslant \phi_{\text{um}}$, $|\epsilon_u(x)| \leqslant \epsilon_{\text{um}}$, $\forall x \in \Omega$. In order to get ϵ_u small we also assume that we have a large number of basis sets, i.e. $N_2 \to \infty$. □

Since the ideal weighs $W_u^\star$ are not known, we introduce *actor estimate weights* $\hat{W}_u \in \mathbb{R}^{N_2 \times m}$ to approximate the optimal control in (28) with $\hat{\bar{\mathcal{K}}}(x)$ (i.e. $u(t) := \hat{\bar{\mathcal{K}}}(x)$) as,

$$\hat{\bar{\mathcal{K}}}(x) = \hat{W}_u^T\phi_u(x), \ \forall x. \tag{29}$$

Our goal is then to tune $\hat{W}_u$ such that the following error is minimized

$$E_u(t) = \frac{1}{2}\,\text{trace}\,\{e_u^T(t)e_u(t)\}, \ \forall t, \tag{30}$$

where

$$e_u \coloneqq \hat{W}_u^T\phi_u + \theta\Big(\frac{1}{2}R^{-1}g^T(x)\nabla\phi^T\hat{W}\Big) \in \mathbb{R}^m.$$

is the error between the estimate (29) and a version of (7) in which $V^\star$ is approximated by the critic's estimate (17).

The tuning for the actor NN is obtained by a gradient-descent-like rule as follows:

$$\begin{aligned}\dot{\hat{W}}_u &= -\alpha_u\frac{\partial E_u}{\partial \hat{W}_u} = -\alpha_u\phi_u e_u \\ &= -\alpha_u\phi_u\Big(\hat{W}_u^T\phi_u + \theta\big(\frac{1}{2}R^{-1}g^T(x)\nabla\phi^T\hat{W}\big)\Big)^T,\end{aligned} \tag{31}$$

where $\alpha_u > 0$ is a constant gain that determines the speed of convergence. Defining the weight estimation error for the actor by

$$\tilde{W}_u \coloneqq W_u^\star - \hat{W}_u \in \mathbb{R}^{N_2 \times m}, \tag{32}$$

and after taking into consideration that (7) with (13) is approximated by (29), the error dynamics can be written as

$$\begin{aligned}\dot{\tilde{W}}_u = -\alpha_u\phi_u\phi_u^T\tilde{W}_u &- \alpha_u\phi_u\theta\big(\frac{1}{2}R^{-1}g^T(x)\nabla\phi^T W^\star\big)^T \\ &+ \alpha_u\phi_u\theta\big(\frac{1}{2}R^{-1}g^T(x)\nabla\phi^T\hat{W}\big)^T \\ &- \alpha_u\phi_u\theta\big(\frac{1}{2}R^{-1}g^T(x)\nabla\epsilon\big)^T - \alpha_u\phi_u\epsilon_u.\end{aligned} \tag{33}$$

Remark 3: Note that the third term of (33) is a function of $\hat{W}$ but since this signal appears inside the saturation function $\theta(\cdot)$, this term is always bounded and will be treated appropriately in the stability analysis that follows. □
A pseudocode (with inline comments to provide guidance following after the symbol ▻) that describes the proposed adaptive-optimal control algorithm has the following form,

Algorithm 2: Adaptive-Optimal Control Algorithm with Relaxed PE

```
Start with initial state $x(0)$, random initial weights $\hat{W}_u(0), \hat{W}(0)$ and $i = 1$
procedure
```

```
    Propagate $t, x(t)$ using (1) and $u(t) := \hat{\hat{\mathcal{K}}}(x)$    ▻ $\{x(t)$ comes from integrating the nonlinear system (1) using any ordinary differential equation (ode) solver (e.g. Runge Kutta) while the time $t$ comes from the Runge Kutta integration process, i.e. $[t_i, t_{i+1}],\ i \in \mathcal{N}$ where $t_{i+1} := t_i + h$ with $h \in \mathbb{R}^+$ the step size$\}$
    Propagate $\hat{W}_u(t), \hat{W}(t)$    ▻ $\{$integrate $\dot{\hat{W}}_u$ as in (31) and $\dot{\hat{W}}$ as in (20) using any ode solver (e.g. Runge Kutta)$\}$
    Compute $\hat{V}(x) = \hat{W}^T\phi(x)$    ▻ output of the Critic NN,
    Compute $\hat{\hat{\mathcal{K}}}(x) = \hat{W}_u^T\phi_u(x)$    ▻ output of the Actor NN
    if $i \neq k$ then    ▻ $\{\{\omega(t_1), \omega(t_2), \dots, \omega(t_i)\}$ has $N$ linearly independent elements and $t_k$ is the time instant that this happens$\}$
        Select an arbitrary data point to be included in the history stack (c.f. Remarks 1-2)
        $i := i + 1$
    end if    ▻ when the history stack is full
end procedure
```

Remark 4: Note that the algorithm runs in real time in a plug-n-play framework and we do not have any iterations. Everything happens simultaneously as we receive new state measurements along the trajectories. One measures the state $x(t)$ and integrates the tuning laws (20) and (31) by using any ordinary differential equation (ode) solver (e.g. Runge Kutta) and then compute $\hat{V}(x) = \hat{W}^T\phi(x)$ and $\hat{\hat{\mathcal{K}}}(x) = \hat{W}_u^T\phi_u(x)$. Numerical methods implemented in modern software packages are mostly adaptive algorithms where, at each step, the step size h is adjusted based on an estimate of the error at that step. In general as h is decreased the calculation takes longer but is more accurate. However, if h is decreased too much the slight rounding that occurs in the computer (because it cannot represent real numbers exactly) begins to accumulate enough to cause significant errors. For many higher order systems, it is very difficult to make the Euler approximation effective. The explicit Runge Kutta methods for non-stiff problems provide computations that are linear to the size of the problem. For stiff problems more accurate, and more elaborate techniques were developed. □

Remark 5: For the proposed method, the involved computation is dominated by the training algorithm for $\hat{W}$ and $\hat{W}_u$ in order to approximate $\hat{V}(x) = \hat{W}^T\phi(x)$ and $\hat{\hat{\mathcal{K}}}(x) = \hat{W}_u^T\phi_u(x)$ which are all variables of the state. If one does the calculations of the right hand side of (20) and (31) in the order of parentheses then, for the critic one has quadratic growth with the number of basis sets N, for the actor one has linear growth with mN_2 and linear growth with the number of states n. Thus the complexity is given as $\mathcal{O}(n + N^2 + mN_2)$ with the term N^2 dominating the other two terms. In order to evaluate the performance of the implemented algorithm, we note that the computational complexity is similar to LSTD [34] (e.g. $\mathcal{O}(n^2)$) but worst than DHP, HDP [35] and Temporal Difference (TD) learning [48] that all have linear complexity with respect to the total number of parameters (e.g. $\mathcal{O}(n)$). Also, we should mention that instead of selecting arbitrary or equally spaced data points as in our work, one can compute the singular values of the history stack matrix and update the history as in [40]. But these computations are very expensive and will boost the algorithmic complexity. □

C. Stability analysis

The regularity assumption is needed for the stability results presented below.

Assumption 3: The process input function g is uniformly bounded on a set $\Omega \subset \mathbb{R}^n$, i.e. $\|g(x)\| < 1/2,\ \forall x \in \Omega$. □

To remove the effect of the NN approximation errors ϵ, ϵ_u (and their partial derivatives) and obtain a closed-loop system with an asymptotically stable equilibrium point, one needs to add a robustifying term to the control law (29) following the work of [49] and use:

$$\hat{\hat{\mathcal{K}}}(x) = \hat{W}_u^T\phi_u(x) + \eta,\ \forall x, \tag{34}$$

where

$$\eta := -B\left\|x\right\|^2 \frac{\mathbf{1}_m}{\left(A + x^Tx\right)},\ \forall x, \tag{35}$$

with A a positive constant, $B \in \mathbb{R}^+$ satisfies $\forall x \in \Omega$

$$\begin{aligned} B\left\|x\right\|^2 \geqslant \frac{A + x^Tx}{(W_\mathrm{m}\phi_\mathrm{dm} + \epsilon_\mathrm{dm})}\bigg\{ & \frac{1}{2\alpha}\bigg(\Big(\frac{k+1}{2}\Big)\epsilon_{Hm}\bigg)^2 \\ & + \left(\phi_\mathrm{um}\bar{u} + \phi_\mathrm{um}\epsilon_\mathrm{um}\right)^2 + \frac{\phi_\mathrm{um}}{2}\left(W_\mathrm{m}\phi_\mathrm{dm} + \epsilon_\mathrm{dm}\right)^2 \\ & + 2(2\phi_\mathrm{um}\bar{u})^2 + (W_\mathrm{m}\phi_\mathrm{dm} + \epsilon_\mathrm{dm})^2 + \epsilon_\mathrm{um}^2\bigg\}. \end{aligned} \tag{36}$$

The following theorem is the main result of the paper and proves asymptotic stability of the learning algorithm of the resulting closed-loop dynamics (1), (35):

$$\dot{x} = f(x) + g(x)\big((W_u^\star - \tilde{W}_u)^T\phi_u(x) + \eta\big), \tag{37}$$

Theorem 3: Consider the closed-loop dynamics given by (37) together with the tuning laws for the critic and the actor NNs given by (20) and (31), respectively. Suppose that the HJB equation (8) has a positive definite, smooth solution, the Assumptions 1, 2, and 3 hold and that $\{\omega(t_1), \omega(t_2), \dots, \omega(t_k)\}$ has N *linearly independent elements*. Then, there exists a triple $\left(\Omega_x \times \Omega_W \times \Omega_{W_u}\right) \subset \Omega$ with Ω compact such that the solution $\tilde{Z} := \big(x(t), \tilde{W}(t), \tilde{W}_u(t)\big) \in \left(\Omega_x \times \Omega_W \times \Omega_{W_u}\right)$ exists globally and converges *asymptotically* to zero for all neural network weights $\tilde{W}(0)$ inside Ω_W, $\tilde{W}_u(0)$ inside Ω_{W_u} and state $x(0)$ inside Ω_x, provided that the following inequalities are satisfied,

$$\alpha > \sqrt{\frac{1}{8\lambda_\mathrm{min}\left(\sum_{i=1}^{k}\frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i)+1)^2}\right)}} \tag{38}$$

$$\phi_\mathrm{um} > \frac{1 + \sqrt{65}}{8}. \tag{39}$$

When the set Ω that appears in the Assumptions 1, 2, and 3 is the whole $\mathbb{R}^n$, then the triple $\Omega_W \times \Omega_{W_u} \times \Omega_x$ can also be the whole $\mathbb{R}^n$. □

Remark 6: For the inequality (38) to hold, one needs to pick the tuning gain α for the critic NN sufficiently large. But as noted in adaptive control [38], large adaptive gains can cause high frequency oscillations in the control signal and reduced tolerance to time delays that will destabilize the system. There are not any systematic approaches to pick a satisfactory adaptation gain, hence trial and error, intuition or Monte Carlo simulations can serve as guidelines. Regarding (39), since ϕ_{um} is simply an upper bound that appears in Assumption 2, one can have it as large as needed. However, one must keep in mind that a large value for ϕ_{um}, requires an appropriate value for the function B (see (36)) in the robustness term in (35). This dependence is clear, from the quadratic terms of ϕ_{um} in (36) (see $\Big\{ \frac{1}{2\alpha}\Big(\Big(\frac{k+1}{2}\Big)\epsilon_{Hm}\Big)^2 + \big(\phi_{\mathrm{um}}\bar{u}+\phi_{\mathrm{um}}\epsilon_{\mathrm{um}}\big)^2+\frac{\phi_{\mathrm{um}}}{2}\big(W_{\mathrm{m}}\phi_{\mathrm{dm}}+\epsilon_{\mathrm{dm}}\big)^2 \quad +2(2\phi_{\mathrm{um}}\bar{u})^2+ (W_{\mathrm{m}}\phi_{\mathrm{dm}}+\epsilon_{\mathrm{dm}})^2+\epsilon_{\mathrm{um}}^2\Big\}$) that increase when one picks larger values for ϕ_{um}. □

Remark 7: By denoting as $\tilde{Z} \coloneqq [\; x^T \quad \tilde{W}^T \quad \tilde{W}_u^T \;]^T$ from the conclusion of Theorem 3 we have that $\left\|\tilde{Z}\right\| \to 0$ which implies $\|x\| \to 0$, it is straightforward that as $t \to \infty$ then from (34) we have (29). □

Remark 8: In case the approximation holds over the entire space, i.e. $\Omega_x = \mathbb{R}^n$, one can conclude global existence of solution provided that the HJB solution $V^\star$ is norm coercive (i.e., $V^\star(x) \to 0 \quad \Rightarrow \quad x \to 0$), as this suffices to guarantee that the Lyapunov function $\mathcal{V}$ that we use in the proof of Theorem 2 is also norm coercive (see [11]). □

IV. Proof of Theorem 3

Consider the following Lyapunov function

$$\mathcal{V} \coloneqq V^\star + \tilde{W}^T\tilde{W} + \frac{1}{2\alpha_u}\operatorname{trace}\{\tilde{W}_u^T\tilde{W}_u\}, \qquad (40)$$

where $V^\star$ is the optimal value function in (5) that is the positive definite and smooth solution of (8) (see Theorem 1), and $V_c := \tilde{W}^T\tilde{W}$ is the Lyapunov function considered in Theorem 2. Since $\mathcal{V}$ is positive definite, there exist class-$\mathcal{K}$ functions $\gamma_1(.)$ and $\gamma_2(.)$ then,

$$\gamma_1(\|\tilde{Z}\|) \leqslant \mathcal{V} \leqslant \gamma_2(\|\tilde{Z}\|),$$

for all $\tilde{Z} \coloneqq [\; x^T \quad \tilde{W}^T \quad \tilde{W}_u^T \;]^T \in B_r$ where $B_r \subset \Omega$ is a ball of radius $r \in \mathbb{R}^+$. By taking the time derivative of the first term with respect to the state trajectories with $u(t)$ (see, (37)) and the second term with respect to the perturbed critic estimation error dynamics (23), using (25), substituting the update for the actor (31) and grouping terms together, then (40) becomes,

$$\begin{aligned}\dot{\mathcal{V}} = \nabla V^\star(x)^T\big(&f(x) - g(x)\tilde{W}_u^T\phi_u \\ &+ g(x)(\bar{\mathcal{K}}^\star(x) - \epsilon_u) - g(x)B\|x\|^2\frac{\mathbf{1}_m}{(A+x^Tx)}\big) \\ &- \frac{\partial V_c}{\partial \tilde{W}}^T\Bigg(\frac{\omega(t)\omega(t)^T}{(\omega(t)^T\omega(t)+1)^2} + \sum_{i=1}^{k}\frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i)+1)^2}\Bigg)\tilde{W} \\ &+ \frac{\partial V_c}{\partial \tilde{W}}^T\Bigg(\frac{\omega(t)}{(\omega(t)^T\omega(t)+1)^2}\epsilon_H(t)\end{aligned}$$

$$\begin{aligned}&+ \sum_{i=1}^{k}\frac{\omega(t_i)}{(\omega(t_i)^T\omega(t_i)+1)^2}\epsilon_H(t_i)\Bigg) \\ &+ \operatorname{trace}\{\tilde{W}_u^T\big(-\phi_u\phi_u^T\tilde{W}_u - \phi_u\theta\Big(\frac{1}{2}R^{-1}g^T(x)\nabla\phi^T\tilde{W}\Big)^T \\ &- \phi_u\theta\Big(\frac{1}{2}R^{-1}g^T(x)\nabla\epsilon\Big)^T - \phi_u\epsilon_u)\},\; t \geqslant 0, \\ &= T_1 + T_2 + T_3. \qquad (41)\end{aligned}$$

where the three terms T_1, T_2, and T_3 are given by equations (42), (43) and (44), respectively.

Using the HJB equation,

$$\nabla V^\star(x)^Tf(x) = -\nabla V^\star(x)^Tg(x)\bar{\mathcal{K}}^\star(x) - R_s(\bar{\mathcal{K}}^\star(x)) - Q(x), \forall x$$

in (44) yields,

$$\begin{aligned}T_3 = &-R_s(\bar{\mathcal{K}}^\star(x)) - Q(x) - \nabla V^\star(x)^Tg(x)\tilde{W}_u^T\phi_u - \nabla V^{\star T}g(x)\epsilon_u \\ &-\nabla V^{\star T}(x)g(x)B\|x\|^2\frac{\mathbf{1}_m}{(A+x^Tx)} \leqslant -R_s(\bar{\mathcal{K}}^\star(x)) - Q(x) \\ &- (W_{\mathrm{m}}\phi_{\mathrm{dm}} + \epsilon_{\mathrm{dm}})\Bigg(\frac{1}{2}\phi_{\mathrm{um}}\|\tilde{W}_u\| \\ &+ (\epsilon_{\mathrm{um}} + \frac{1}{2}B\|x\|^2\frac{\mathbf{1}_m}{A+x^Tx})\Bigg) \qquad (45)\end{aligned}$$

since $A + x^Tx > 0$. The term T_3 can be further upper bounded as,

$$\begin{aligned}T_3 \leqslant &-R_s(\bar{\mathcal{K}}^\star(x)) - Q(x) + \frac{\phi_{\mathrm{um}}}{4}\big((W_{\mathrm{m}}\phi_{\mathrm{dm}} + \epsilon_{\mathrm{dm}})\big)^2 \\ &+ \frac{\phi_{\mathrm{um}}}{4}\|\tilde{W}_u\|^2 + \frac{1}{2}(W_{\mathrm{m}}\phi_{\mathrm{dm}}+ \\ &\epsilon_{\mathrm{dm}})^2 + \frac{1}{2}\epsilon_{\mathrm{um}}^2 - (W_{\mathrm{m}}\phi_{\mathrm{dm}} + \epsilon_{\mathrm{dm}})\frac{1}{2}B\|x\|^2\frac{\mathbf{1}_m}{A+x^Tx}. \qquad (46)\end{aligned}$$

Finally after taking into account the bound of $B\|x\|^2$ from (36) we can upper bound (41) as,

$$\begin{aligned}\dot{\mathcal{V}} \leqslant &-(2\alpha\lambda_{\min}(\sum_{i=1}^{k}\frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i)+1)^2}) - \frac{1}{4\alpha})\|\tilde{W}\|^2 \\ &- (\phi_{\mathrm{um}}^2 - \frac{\phi_{\mathrm{um}}}{4} - 1)\|\tilde{W}_u\|^2 \\ &- R_s(\bar{\mathcal{K}}^\star(x)) - Q(x),\; t \geqslant 0. \qquad (47)\end{aligned}$$

Then, by taking into account the inequalities (38) and (39) one has, $\dot{\mathcal{V}} \leqslant 0, \;\; t \geqslant 0$. From Barbalat's lemma [50] it follows that as $t \to \infty$, then $\|\tilde{Z}\| \to 0$.

The result holds as long as we can show that the state $x(t)$ remains in the set $\Omega \subseteq \mathbb{R}^n$ for all times. To this effect, define the following compact set

$$M \coloneqq \{x \in \mathbb{R}^n | \mathcal{V}(t) \leqslant m)\} \subset \mathbb{R}^n$$

where m is chosen as the largest constant so that $M \subseteq \Omega$. Since by assumption $x_0 \in \Omega_x$, and $\Omega_x \subset \Omega$ then we can conclude that $x_0 \in \Omega$. While $x(t)$ remains inside Ω, we have seen that $\dot{\mathcal{V}} \leqslant 0$ and therefore $x(t)$ must remain inside $M \subset \Omega$. The fact that $x(t)$ remains inside a compact set also excludes the possibility of finite escape time and therefore one has global existence of solution. ■

$$
\begin{aligned}
T_1 \coloneqq {} & -\frac{\partial V_c}{\partial \tilde{W}}^T \left(\frac{\omega(t)\omega(t)^T}{(\omega(t)^T\omega(t)+1)^2} + \sum_{i=1}^{k} \frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i)+1)^2} \right) \tilde{W} + \frac{\partial V_c}{\partial \tilde{W}}^T \left(\frac{\omega(t)}{(\omega(t)^T\omega(t)+1)^2} \epsilon_H(t) + \sum_{i=1}^{k} \frac{\omega(t_i)}{(\omega(t_i)^T\omega(t_i)+1)^2} \epsilon_H(t_i) \right) \\
& \leqslant -2\alpha\lambda_{\min}\big(\sum_{i=1}^{k} \frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i)+1)^2} \big) \|\tilde{W}\|^2 + \frac{1}{2\alpha} \|\tilde{W}\| \left((\frac{k+1}{2})\epsilon_{Hm} \right) \\
& \leqslant -2\alpha\lambda_{\min}\big(\sum_{i=1}^{k} \frac{\omega(t_i)\omega(t_i)^T}{(\omega(t_i)^T\omega(t_i)+1)^2} \big) \|\tilde{W}\|^2 + \frac{1}{4\alpha} \|\tilde{W}\|^2 + \frac{1}{4\alpha} \left((\frac{k+1}{2})\epsilon_{Hm} \right)^2 \qquad (42)
\end{aligned}
$$

$$
\begin{aligned}
T_2 \coloneqq {} & \operatorname{trace}\{ \tilde{W}_u^T \big(-\phi_u \phi_u^T \tilde{W}_u - \phi_u \theta \left(\frac{1}{2} R^{-1} g^T(x) \nabla\phi^T W^\star \right)^T \\
& + \phi_u \theta \left(\frac{1}{2} R^{-1} g^T(x) \nabla\phi^T \hat{W} \right)^T - \phi_u \theta \left(\frac{1}{2} R^{-1} g^T(x) \nabla\epsilon \right)^T - \phi_u \epsilon_u \big) \} \\
& \leqslant -\phi_{\mathrm{um}}^2 \|\tilde{W}_u\|^2 + 2\phi_{\mathrm{um}}\bar{u}\|\tilde{W}_u\| + \big(\phi_{\mathrm{um}}\bar{u} + \phi_{\mathrm{um}}\epsilon_{\mathrm{um}}\big)\|\tilde{W}_u\| \\
& \leqslant -\phi_{\mathrm{um}}^2 \|\tilde{W}_u\|^2 + (2\phi_{\mathrm{um}}\bar{u})^2 + \frac{\big(\phi_{\mathrm{um}}\bar{u} + \phi_{\mathrm{um}}\epsilon_{\mathrm{um}}\big)^2}{2} + \|\tilde{W}_u\|^2. \qquad (43)
\end{aligned}
$$

$$
T_3 \coloneqq \nabla V^\star(x)^T \big(f(x) - g(x)\tilde{W}_u^T \phi_u + g(x)(u^\star(x) - \epsilon_u) - g(x) B \, \|x\|^2 \, \frac{\mathbf{1}_m}{\big(A + x^T x\big)} \big) \qquad (44)
$$

V. Simulations

This section presents two simulation examples to illustrate the effectiveness of the proposed optimal adaptive control algorithm. In the simulations below, since the history stack is empty in the beginning we need to add a dithering noise to the control input in the form of $\rho(t) = \frac{1}{2}\big(sin(0.3\pi t) + cos(0.3\pi t)\big)$ for the first second.

A. Van-der Pol Oscillator

Consider a controlled Van-der Pol oscillator of the form (1), with

$$
f(x) = \left[\begin{smallmatrix} x_2 \\ -x_1 - \frac{1}{2}x_2(1-x_1^2) - x_1^2 x_2 \end{smallmatrix} \right], \quad g(x) = \left[\begin{smallmatrix} 0 \\ x_1 \end{smallmatrix} \right], \qquad (48)
$$

which has an uncontrolled unstable limit cycle and a stable uncontrollable equilibrium point at the origin. It is shown in [51] that the non-saturated optimal control input, with a criterion $V(x(0)) = \int_0^\infty (\|x\|^2 + \|u\|^2) d\tau$, is given by $\bar{\mathcal{K}}^\star(x) = -x_1 x_2$ and that the corresponding optimal value function is given by $V^\star(x) = x_1^2 + x_2^2$.

We now consider the optimal control of (48) with the input saturated so that $|u| \leqslant \bar{u} = 0.1$ and the cost defined by (2), (3), and (4) with $Q(x) \coloneqq \|x\|^2$ and $R = 1$; for which the optimal feedback law is not known in closed form. The NN weights are initialized randomly in the interval $[0, 1]$, the activation functions were chosen to be quadratics of the form $\phi(x) = [x_1^2 \; x_1 x_2 \; x_2^2]^T$ and $\phi_u(x) \coloneqq \nabla\phi(x)$, and the tuning gains were set to $\alpha = 10, \alpha_u = 2$. Thus, the critic parameters converged to $\hat{W} = \begin{bmatrix} 2.4953 & 0.9991 & 2.2225 \end{bmatrix}$. Figure 1 presents the phase plane trajectory of the system and the optimal control input, which is saturated when it reaches the maximum and minimum saturation limits. Figure 2 shows the convergence of the critic parameters which takes almost 5 seconds to converge.

It is well known that parameter convergence cannot be achieved for nonlinear systems without PE. Since PE is unverifiable for general nonlinear systems, trying to achieve parameter convergence for such systems is very difficult. In order to show the efficacy of our proposed approach with relaxed PE compared to the tuning law (20) without the second term (i.e. past data), we will compare the result from figure 2 to two different cases. The first case considers a "strong PE" (i.e. a large number of sinusoids of different frequencies and high amplitude) that is applied for $0 \leqslant t \leqslant 40$, and is shown in figure 3. The second case of a "weak PE" (i.e. a modest number of sinusoids of different frequencies and low amplitude) that is applied for $0 \leqslant t \leqslant 20$ seconds, is shown in figure 4. From figures 3-4, one shall see how difficult is to guarantee PE throughout learning for every $t \geqslant 0$. Specifically, in the first case the weight estimates reach the optimal solution after 20 seconds (compared to just 5 seconds with the proposed algorithm) and in the second case the weights converge fast but get stuck in a local minimum. This happens since the PE condition given in definition 2 is violated due to the reason that the states x reach zero either too late (i.e. $\omega(t)$ becomes zero after oscillating) or too early (i.e. $\omega(t)$ becomes zero before the weights reach the optimal solution). Our proposed learning framework with previous data solves these issues.

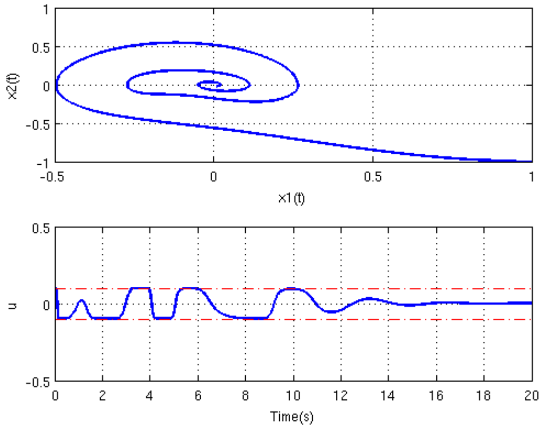

Fig. 1. The top plot shows the phase plane trajectory of the closed-loop system shows convergence to the origin. The bottomplots shows the control input, which is saturated when it reaches the saturation limits.

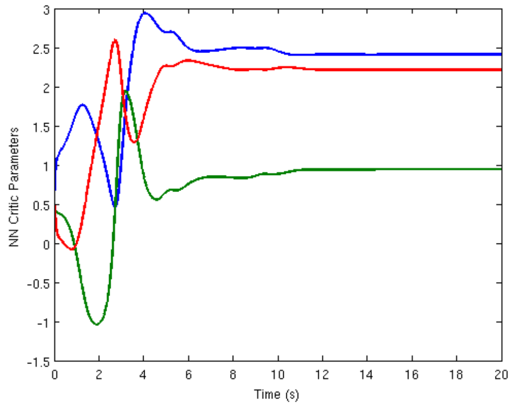

Fig. 2. Convergence of the critic parameters to the optimal cost.

B. Power Plant System

Consider the power system shown in Figure 5, consisting of a turbine-generator, a system load, and an automatic generation control. A simplified state-space model for this system is of the form

$$\begin{bmatrix} \Delta\dot{\bar{\alpha}} \\ \Delta\dot{P}_m \\ \Delta\dot{f}_G \end{bmatrix} = \begin{bmatrix} -\frac{1}{T_g} & 0 & \frac{1}{R_g T_g} \\ \frac{K_t}{T_t} & -\frac{1}{T_t} & 0 \\ 0 & \frac{K_p}{T_p} & -\frac{1}{T_p} \end{bmatrix} \begin{bmatrix} \Delta\bar{\alpha} \\ \Delta P_m \\ \Delta f_G \end{bmatrix} + \begin{bmatrix} \frac{1}{T_g} \\ 0 \\ 0 \end{bmatrix} \Delta P_c, \quad (49)$$

where Δf_G is the incremental frequency deviation, ΔP_m is the incremental change in the generator output, $\Delta\bar{\alpha}$ is the incremental change in governor value position, and the control input ΔP_c of the system is the incremental speed change in position deviation [52]. The system parameters include the governor time constant $T_g = 0.08[s]$, the turbine time constant $T_t = 0.1[s]$, the generator model time constant $T_p = 20[s]$, the feedback regulation constant $R_g = 2.5[Hz/MW]$, the generator model gain constant $K_p = 120[Hz/MW]$, and the turbine model gain constant $K_t = 1[s]$.

The control objective is to keep the frequency of generator f_G, the governor valve position $\bar{\alpha}$, and the generator output P_m in their nominal values, despite perturbations in the load. To this effect, we considered the optimal control of (49) with the input saturated so that $|\Delta P_c| \leqslant \bar{u} = 0.02$ and the cost defined by (2), (3), and (4) with $Q(x) \coloneqq \Delta\bar{\alpha}^2 + \Delta P_m^2 + \Delta f_G^2$ and $R = 0.5$. The initial NN weights were randomly initialized in $[0, 1]$,

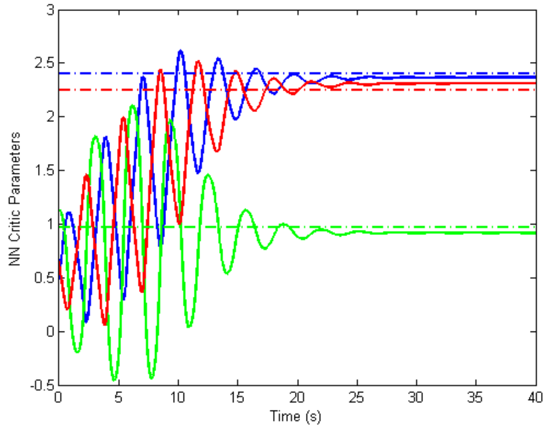

Fig. 3. Evolution of the critic parameters with a "strong" persistence of excitation applied for $0 \leqslant t \leqslant 40$ seconds. With the dashed lines one can observe the optimal critic parameters as plotted in Figure 2.

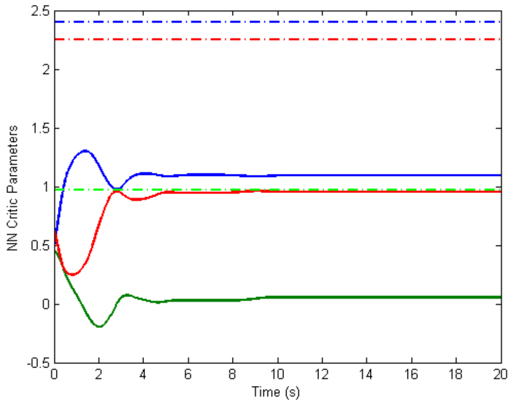

Fig. 4. Evolution of the critic parameters with a "weak" persistence of excitation applied for $0 \leqslant t \leqslant 20$ seconds. With the dashed lines one can observe the optimal critic parameters as plotted in Figure 2.

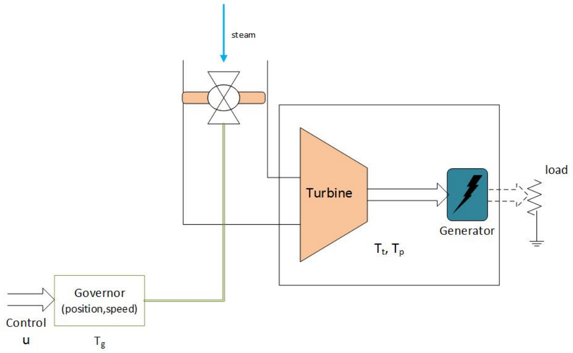

Fig. 5. Power system block diagram.

tuning gains set to $\alpha = 10$, $\alpha_u = 2$, and the activation functions were chosen to be quadratics of the form $\phi(x) = [x_1^2\ x_1x_2\ x_1x_3\ x_2^2\ x_2x_3\ x_3^2]^T$, $\phi_u(x) \coloneqq \nabla\phi(x)$. The critic neural network weights converged to $\hat{W} = \big[0.0583 \quad 0.0476 \quad 0.0549 \quad 0.1123 \quad 0.1447 \quad 0.3489\big]$. Figure 6 shows the state evolution, Figure 7 the incremental speed change in position deviation (control input), and Figure 8 the evolution of the critic NN weights. A perturbation of 5% is applied to the generator frequency in the interval 7-11 seconds and we can see the system's adaptation to the new load.

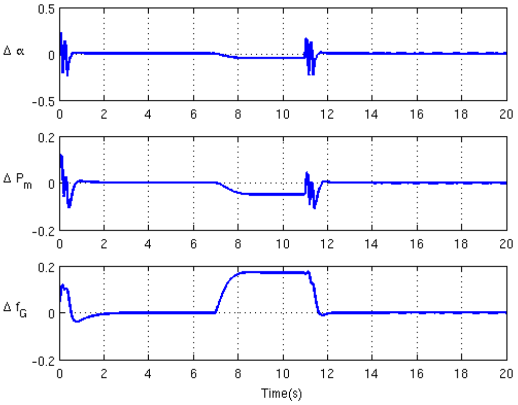

Fig. 6. Time evolution of the power system states with a perturbation of 5% applied in the generator frequency during $7s - 11s$.

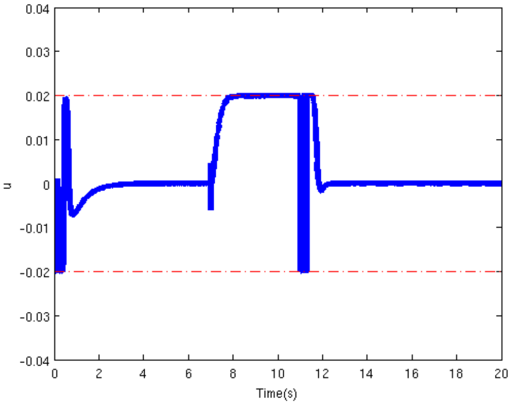

Fig. 7. Incremental speed change in position deviation (control input of the system).

VI. Conclusion

This paper proposed a new approximate dynamic programming algorithm for systems with bounded inputs, which relaxes the persistence of excitation condition by using previously stored data concurrently with current data. To suppress the effects of the critic and actor NN approximation errors, a new robustifying term was added to the controller. By considering an appropriate Lyapunov function, we prove asymptotic stability of the overall closed-loop system. Simulation results of a controlled Van-der Pol oscillator and a power system illustrate the effectiveness and efficiency of the proposed approach. Future work will be concentrated on extending the results for completely unknown systems and multiple decision makers.

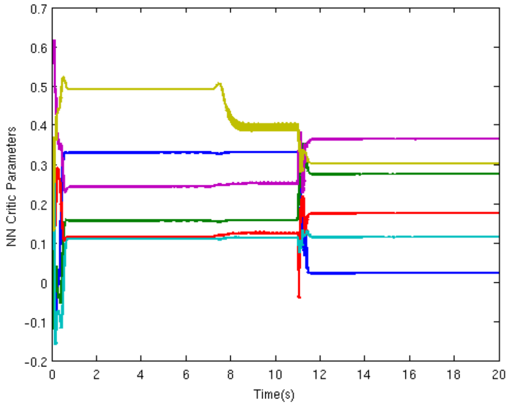

Fig. 8. Convergence of the critic parameters to the optimal cost.

References

[1] P. J. Werbos, "Brain-like intelligent control: from neural nets to larger-scale systems," in *ESANN*, pp. 59–66, 1998.

[2] P. J. Werbos, "Intelligence in the brain: A theory of how it works and how to build it," *Neural Networks*, vol. 22, no. 3, pp. 200 – 212, 2009.

[3] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 3rd ed., 2007.

[4] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 32, no. 2, pp. 140–153, 2002.

[5] W. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley Series in Probability and Statistics, Wiley, 2007.

[6] S. Bhasin, R. Kamalapurkar, M. Johnson, K. Vamvoudakis, F. Lewis, and W. Dixon, "A novel actor critic identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82 – 92, 2013.

[7] F. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. IEEE Press Series on Computational Intelligence, Wiley, 2013.

[8] F. Lewis, D. Vrabie, and K. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 76 –105, 2012.

[9] K. G. Vamvoudakis and F. L. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[10] Q.-L. Wei, H.-G. Zhang, and L.-L. Cui, "Data-based

optimal control for discrete-time zero-sum games of 2-d systems using adaptive critic designs," *Acta Automatica Sinica*, vol. 35, no. 6, pp. 682 – 692, 2009.

[11] H. K. Khalil, *Nonlinear systems*. Macmillan Pub. Co., 1992.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press, 1998.

[13] S. E. Lyshevski, "Optimal control of nonlinear continuous-time systems: design of bounded controllers via generalized non-quadratic functionals," in *Proc. American Control Conference*, vol. 1, pp. 205 –209 vol.1, 1998.

[14] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[15] H. Zhang, C. Qin, and Y. Luo, "Neural-network-based constrained optimal control scheme for discrete-time switched nonlinear system using dual heuristic programming," *Automation Science and Engineering, IEEE Transactions on*, vol. 11, pp. 839–849, July 2014.

[16] Q. Yang and S. Jagannathan, "Reinforcement learning controller design for affine nonlinear discrete-time systems using online approximators," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 377 –390, 2012.

[17] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming for nonlinear control design," in *IEEE 51st Conference on Decision and Control (CDC)*, pp. 1896 –1901, 2012.

[18] R. Song, W. Xiao, H. Zhang, and C. Sun, "Adaptive dynamic programming for a class of complex-valued nonlinear systems," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, pp. 1733–1739, Sept 2014.

[19] J. Liang, G. K. Venayagamoorthy, and R. G. Harley, "Wide-area measurement based dynamic stochastic optimal power flow control for smart grids with high variability and uncertainty," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 59 –69, 2012.

[20] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[21] S. Adam, L. Busoniu, and R. Babuska, "Experience replay for real-time reinforcement learning control," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 2, pp. 201–212, 2012.

[22] P. Cichosz, "An analysis of experience replay in temporal difference learning," *Cybernetics & Systems*, vol. 30, no. 5, pp. 341–363, 1999.

[23] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-based batch mode reinforcement learning," in *Journal of Machine Learning Research*, pp. 503–556, 2005.

[24] M. Riedmiller, "Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method," in *Machine Learning: ECML 2005*, pp. 317–328, Springer, 2005.

[25] M. Fairbank and E. Alonso, "A comparison of learning speed and ability to cope without exploration between DHP and TD(0)," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pp. 1–8, June 2012.

[26] P. J. Werbos, "Stable adaptive control using new critic designs," in *Ninth Workshop on Virtual Intelligence/Dynamic Neural Networks*, pp. 510–579, International Society for Optics and Photonics, 1999.

[27] L. Baird *et al.*, "Residual algorithms: Reinforcement learning with function approximation," in *ICML*, pp. 30–37, 1995.

[28] P. Werbos, A. G. Barto, and T. G. Dietterich, "ADP: Goals, opportunities and principles," *Handbook of learning and approximate dynamic programming*, vol. 2, p. 1, 2004.

[29] D. V. Prokhorov and D. C. Wunsch, "Convergence of critic-based training," in *IEEE International Conference on Systems Man and Cybernetics*, vol. 4, pp. 3057–3060, Citeseer, 1997.

[30] P. J. Werbos, "Consistency of HDP applied to a simple reinforcement learning problem," *Neural Networks*, vol. 3, no. 2, pp. 179 – 189, 1990.

[31] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, "Fast gradient-descent methods for temporal-difference learning with linear function approximation," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML, (New York, NY, USA), pp. 993–1000, ACM, 2009.

[32] H. R. Maei, C. Szepesvári, S. Bhatnagar, D. Precup, D. Silver, and R. S. Sutton, "Convergent temporal-difference learning with arbitrary smooth function approximation.," in *NIPS*, pp. 1204–1212, 2009.

[33] M. Fairbank, E. Alonso, and D. Prokhorov, "An equivalence between adaptive dynamic programming with a critic and backpropagation through time," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, pp. 2088–2100, Dec 2013.

[34] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine Learning*, vol. 22, no. 1-3, pp. 33–57, 1996.

[35] P. J. Werbos, "A menu of designs for reinforcement learning over time," *Neural networks for control*, pp. 67–95, 1990.

[36] Q. Wei, F.-Y. Wang, D. Liu, and X. Yang, "Finite-approximation-error-based discrete-time iterative adaptive dynamic programming," *Cybernetics, IEEE Transactions on*, vol. 44, pp. 2820–2833, Dec 2014.

[37] P. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of Intelligent Control* (White and Sofge, eds.), New York: New York: Van Nostrand Reinhold, 1992.

[38] P. Ioannou and B. Fidan, *Adaptive Control Tutorial*. Advances in Design and Control, Society for Industrial and Applied Mathematics, 2006.

[39] Z. Chen and S. Jagannathan, "Generalized HJB formu-

lation: Based neural network control of affine nonlinear discrete-time systems," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 90–106, 2008.

[40] G. Chowdhary and E. Johnson, "Concurrent learning for convergence in adaptive control without persistency of excitation," in *IEEE Conference on Decision and Control (CDC)*, pp. 3674 –3679, 2010.

[41] J. A. Boyan, "Technical update: Least-squares temporal difference learning," *Machine Learning*, vol. 49, no. 2-3, pp. 233–246, 2002.

[42] A. Heydari and S. N. Balakrishnan, "Fixed-final-time optimal control of nonlinear systems with terminal constraints," *Neural Netw.*, vol. 48, pp. 61–71, Dec. 2013.

[43] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193 – 202, 2014.

[44] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114 – 115, 1968.

[45] H. W. K. Hornik, M. B. Stinchcombe, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, no. 5, pp. 551–560, 1990.

[46] G. Chowdhary, T. Yucelen, M. Mhlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *International Journal of Adaptive Control and Signal Processing*, 2012.

[47] T. Dierks and S. Jagannathan, "Optimal control of affine nonlinear continuous-time systems," in *American Control Conference (ACC), 2010*, pp. 1568–1573, IEEE, 2010.

[48] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[49] M. Polycarpou, J. Farrell, and M. Sharma, "On-line approximation control of uncertain nonlinear systems: issues with control input saturation," in *Proc. American Control Conference*, vol. 1, pp. 543–548 vol.1, 2003.

[50] W. Haddad and V. Chellaboina, *Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach*. Princeton University Press, 2008.

[51] V. Nevistic and J. A. Primbs, "Constrained nonlinear optimal control: A converse HJB approach," tech. rep., 1996.

[52] D. Iracleous and A. Alexandridis, "A multi-task automatic generation control for power regulation," *Electric Power Systems Research*, vol. 73, no. 3, pp. 275 – 285, 2005.