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

Kyriakos G. Vamvoudakis\textsuperscript{1}, Member, IEEE \quad Marcio F. Miranda\textsuperscript{2} \quad João P. Hespanha\textsuperscript{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\cite{1,2} and bring together the advantages of adaptive and optimal control to obtain approximate and forward in time solutions to difficult optimization problems\cite{3–5}. But all the existing algorithms — such as the ones developed in\cite{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\cite{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\cite{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)\cite{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\cite{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\cite{14} used neural networks to solve the HJB offline. The work of\cite{15} 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\cite{16} 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. Several adaptive/critic designs have been applied in real applications, such as\cite{17}, 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\cite{12,18} incrementally update the policies based on each individual experience. The combination of experience replay\cite{19,20} and fitted Q iteration\cite{21,22}, 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 [23, 24] have shown that dual heuristic dynamic programming (DHDP) 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 [24] states that for any given plant, data from optimal and suboptimal action strategies can be used without difficulty in system identification which means that DHDP has less excitation problems. There is an extensive research work on linear discrete-time systems with “residual gradients” [25] or “Galerkinized” methods [24, 26, 27], where the critic NN weights do not converge to the optimal values in the stochastic case [28, 29], but convergence is guaranteed in deterministic case. The aforementioned limitations of “Galerkinized” methods have been overcome in [30] where the authors solved this problem under standard technical assumptions. In [31] the authors propose a variant of dual heuristic dynamic programming that has a guaranteed convergence, under certain smoothness conditions and a greedy policy, when using a general smooth nonlinear function approximator for the critic. Our work in this paper is focused on a “Galerkinized” asymptotically stable algorithm for known deterministic nonlinear systems.

There is a lot of research work on approximate dynamic programming for linear systems [24, 32] where the persistence of excitation condition used also in adaptive control [33] needs to be satisfied in order to guarantee convergence of the critic NN. For nonlinear systems, there is no clear recent development. Adaptive optimal control algorithms with approximate dynamic programming [6, 8, 34] 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 [35] from the adaptive control side, and the works of [36] and [37] 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. The authors in [38] have used concurrent learning in optimal adaptive control but they only proved 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 \( * \) is used to denote the optimal solution of an optimization, \( \lambda_{\text{min}}(A) \) 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 \( \nabla \). A function \( \alpha : \mathbb{R}^+ \to \mathbb{R} \) is said to belong to class \( K(\alpha \in K) \) functions if it is continuous, strictly increasing and \( \alpha(0) = 0 \).

II. PROBLEM FORMULATION

Consider the nonlinear continuous-time system given by
\[
\dot{x}(t) = f(x(t)) + g(x(t))u(t); \quad x(0) = x_0, \quad t \geq 0 \tag{1}
\]
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 such that \( x \mapsto f(x) + g(x)u(x) \) is locally Lipschitz continuous in \( x \) for fixed \( U \) so that the solution \( x(t) \) exists and is unique for any initial condition \( x_0 \) and piecewise continuous control \( u(t) \in U \). We assume that the \( x(t) \) is available for full state feedback.

It is desired to minimize the following infinite horizon cost functional
\[
V(x(0)) = \int_0^\infty r(x(t), u(t)) \, dt, \quad \forall x(0) \tag{2}
\]
with
\[
r(x, u) = Q(x) + R_s(u), \quad \forall x, u, u \tag{3}
\]
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 at the origin 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,
\[
\bar{x} \in B(x_e, \delta) \Rightarrow \left\{ \psi(t; 0, \bar{x}) \in B(x_e, \epsilon), \quad \forall t \geq 0 \right\}
\]
\[
\lim_{t \to \infty} \psi(t; 0, \bar{x}) = x_e,
\]
where $B(\bar{x}, \epsilon)$ denotes the open ball centered at $\bar{x}$ of radius $\epsilon$, i.e. the set, $\{x \in \mathbb{R}^n : |x - \bar{x}| < \epsilon\}$.

The optimal control problem is to find an admissible control $u^*(t)$ 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. $|u_i| \leq \bar{u}, \forall i \in \{1, \ldots, m\}$) we follow the approach in [13] and use a nonquadratic penalty function of the form,

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

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

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

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 $C^n$, $n \geq 1$, 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. E.g., one could select

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

$$:= 2 \int_0^{u} (\bar{u} \tanh^{-1}(v/\bar{u}))^T R dv > 0, \forall u.$$ (4)

The optimal value function is defined as,

$$V^*(x(t)) = \min_{u \in U} \int_t^{\infty} r(x, u) dt, t \geq 0,$$ (5)

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

$$H(x, u, \nabla V) = \nabla V(x)^T \left( f(x) + g(x)u(x) \right) + Q(x) + R_s(u), \forall x, u$$ (6)

The constrained optimal control input for the system (1), with cost (3), (4), (5), can be obtained using the stationarity condition in the Hamiltonian (6):

$$\frac{\partial H(x, u, \nabla V)}{\partial u} = 0$$

$$\Rightarrow \nabla V^*(x)^T g(x) + 2R\theta^{-T}(u^*(x)) = 0$$

$$\Rightarrow u^*(x) = -\theta \left( \frac{1}{2} R^{-1} g(x) \nabla V^*(x)^T \right).$$ (7)

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

$$H^*(x, u^*, \nabla V^*) := \nabla V^*(x)^T \left( f(x) + g(x)u^*(x) \right) + Q(x) + R_s(u^*(x)) = 0, \forall x.$$ (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

$$H(x, u, \nabla V) = 0,$$ (9)

with $H(\cdot)$ given by (8) and $u$ given by

$$u = -\theta \left( \frac{1}{2} R^{-1} g(x) \nabla V(x)^T \right).$$ (10)

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

$$V = \nabla V(x)^T (f(x) + g(x)u) = -R_s(u) - Q(x) < -Q(x)$$

where we used (9) with $u$ 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

$$V^*(x(0)) + \int_0^{x} \nabla V^*(x)^T (f(x) + g(x)u) = 0.$$ (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,

$$V(x(0); u) = \int_0^{x} \left( R_s(u) + Q(x) \right) dt + V^*(x(0))$$

$$+ \int_0^{x} \nabla V^*(x)^T (f(x) + g(x)u) dt.$$ By subtracting zero (using the HJB equation) we have,

$$V(x(0); u) = \int_0^{x} \left( R_s(u) - R_s(u^*) \right) dt + V^*(x(0)).$$ (12)

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

$$V(x(0); u) = \int_0^{x} \left( 2\left( \int_0^{u} (\theta^{-1}(v))^T dv - \int_0^{u^*} (\theta^{-1}(v))^T dv \right) \right)$$

$$- R^T(u^*) dt + V^*(x(0)).$$

We can complete the squares and hence we have,

$$V(x(0); u) = \int_0^{x} R_s(u - u^*) dt + V^*(x(0)).$$

Now by setting $u = u^*$ one can show that,

$$V^*(x(0)) \leq V(x(0); u),$$

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 \( \epsilon_{ac} \) is a small number used to terminate the algorithm when two consecutive value functions differ by less than \( \epsilon_{ac} \). In the linear case, this algorithm reduces to Kleinman’s algorithm [39].

**Algorithm 1: Policy Iteration for Nonlinear Systems**

1: procedure
2:  Given admissible policies \( \mu^{(0)} \)
3:  while \( \|V^{(i)} - V^{(i-1)}\| \geq \epsilon_{ac} \) do
4:     Solve for the value \( V^{(i)}(x) \) using Bellman’s equation
5:       \( Q(x) + \nabla V^{(i)}(x) = \frac{1}{2} R^{-1} g^T(x) \nabla V^{(i)}(x)^T \)
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 \( V^*(x) \) in (5) with a critic neural network (NN), within a set \( \Omega \subset \mathbb{R}^n \) that contains the origin, as follows

\[
V^*(x) = W^T \phi(x) + \epsilon(x), \quad \forall x
\]

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

One should pick the NN activation functions \( \varphi_i(x), \forall i \in \{1, 2, \ldots, N\} \) as quadratic, radial basis or sigmoid functions so that they define a complete independent basis set for \( V^* \).

In this case, \( V^* \) and its derivatives

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

can be uniformly approximated on any given compact set \( \Omega \). According to Weierstrass Higher Order Approximation Theorem [40], [14], as the number of basis sets \( N \) increases, the approximation error on a compact set \( \Omega \) goes to zero, i.e., \( \epsilon(x) \rightarrow 0 \) as \( N \rightarrow \infty \). We shall require a form of uniformity in this approximation result that is common in neuro-adaptive control and other approximation techniques [33, 40]. This assumption also involves the approximate Hamiltonian defined by

\[
H^*(x, u, W^* \nabla \phi) := \hat{W}^* \nabla \phi(x)(f + gu)
\]

which is obtained by using (14) in (6) and that leads to the residual error

\[
\epsilon_H := H^* - H = -\nabla \epsilon^T (f + gu), \quad \forall x, u.
\]

**Assumption 1 (Critic Uniform Approximation):**
The critic activation functions \( \phi \), the value function approximation error \( \epsilon \), their derivatives, and the Hamiltonian residual error \( \epsilon_H \) are all uniformly bounded on a set \( \Omega \subset \mathbb{R}^n \), in the sense that there exist finite constants \( \phi_m, \phi_m, \epsilon_m, \epsilon_m, \epsilon_H \in \mathbb{R}^+ \) such that \( |\phi(x)| \leq \phi_m, |\nabla \phi(x)| \leq \phi_m, |\epsilon(x)| \leq \epsilon_m, |\nabla \epsilon(x)| \leq \epsilon_m, |\epsilon_H(x)| \leq \epsilon_H, \forall x \in \Omega. \) In order to get \( \epsilon \) small we also assume that we have a large number of basis sets, i.e. \( N \rightarrow \infty \).

Since the ideal weights \( W^* \) for the (approximate) value function \( V^*(x) \) 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:

\[
\hat{V}(x) = \hat{W}^T \phi(x), \forall x.
\]

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

\[
\hat{H}(x, u, \hat{W}^T \nabla \phi) := \hat{W}^T \nabla \phi(x)(f + gu) \]

+ \( Q(x) + R_s(u), \quad \forall x, u, \) (15)

for the (approximate) Hamiltonian.

**Definition 2:** [33] A vector signal \( \Phi(t) \) is exciting over the interval \( [t, t + T_{PE}] \), with \( T_{PE} \in \mathbb{R}^+ \) if there exists \( \beta_1, \beta_2 \in \mathbb{R}^+ \) such that \( \beta_1 I \leq \int_t^{t+T} \Phi(\tau) \Phi^T(\tau) d\tau \leq \beta_2 I \) with \( I \) an identity matrix of appropriate dimensions.

To achieve convergence of (18) to the (approximate) Hamiltonian (15) along closed-loop trajectories, one would typically need persistency of excitation (see Definition 2) for the vector \( \omega(t) \) defined by

\[
\omega(t) := \nabla \phi(x(t))(f(x(t)) + g(x(t))u(x(t))\)
\]

along the closed-loop trajectories [33]. 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 [41] 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 \( t \):

\[
\epsilon(t) := \hat{H} \left( x(t), u(t), \hat{W}(t)^T \nabla \phi(x(t)) \right) - H^*(x, u^*, \nabla V^*)
\]

\[
= \hat{H} \left( x(t), u(t), \hat{W}(t)^T \nabla \phi(x(t)) \right)
\]

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

\[
\epsilon_{buff}(t_i, t) := \hat{H} \left( x(t_i), u(t_i), \hat{W}(t)^T \nabla \phi(x(t)) \right)
\]
We draw attention to the reader that, while the error \( e_{\text{buff}}(t, t') \) uses past state and input data \( x(t) \) and \( u(t) \), respectively, it is defined based on the current weight estimates \( W(t) \).

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

\[
E(t) = \frac{1}{2} \left( \frac{e(t)^2}{(\omega^T \omega + 1)^2} + \sum_{i=1}^k \frac{e_{\text{buff}}(t, t)^2}{(\omega(t_i)^T \omega(t_i) + 1)^2} \right), \quad \forall t.
\]

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

\[
\dot{W} = -\alpha \frac{\delta E}{\delta W} = -\alpha \frac{\omega(t)e(t)}{(\omega(t)^T \omega(t) + 1)^2} - \alpha \sum_{i=1}^k \frac{\omega(t_i)e_{\text{buff}}(t, t_i)}{(\omega(t_i)^T \omega(t_i) + 1)^2} 
= -\alpha \frac{\omega(t)(\omega(t)^T \dot{W}(t) + R_u(u(t)) + Q(x(t)))}{(\omega(t)^T \omega(t) + 1)^2} 
- \alpha \sum_{i=1}^k \frac{\omega(t_i)(\omega(t_i)^T \dot{W}(t) + Q(x(t_i)) + R_u(u(t_i)))}{(\omega(t_i)^T \omega(t_i) + 1)^2},
\]

\[
\forall t > t_i \geq 0, \text{ where } \alpha > 0 \text{ is a constant gain that determines the speed of convergence. Defining the weight estimation error of the critic by } \bar{W} := W^* - \dot{W} \in \mathbb{R}^N.
\]

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

\[
\dot{\bar{W}} = -N_{\text{om}} + P_{\text{err}}.
\]

where,

\[
N_{\text{om}}(t) := \alpha \left( \frac{\omega(t)^T \omega(t)}{(\omega(t)^T \omega(t) + 1)^2} + \sum_{i=1}^k \frac{\omega(t_i)^T \omega(t_i)}{(\omega(t_i)^T \omega(t_i) + 1)^2} \right) \bar{W}(t),
\]

(23)

can be viewed as defining the nominal error dynamics and

\[
P_{\text{err}}(t) := \alpha \left( \frac{\omega(t)^T \dot{W}(t) + \epsilon_H(t)}{(\omega(t)^T \omega(t) + 1)^2} + \sum_{i=1}^k \frac{\omega(t_i)^T \dot{W}(t_i) + \epsilon_H(t_i)}{(\omega(t_i)^T \omega(t_i) + 1)^2} \right)
\]

(24)

a perturbation term, bounded as \( |P_{\text{err}}| \leq \frac{\alpha}{2} (k+1) \epsilon_{\text{Hm}}, \) that would be zero if the Hamiltonian errors \( \epsilon_H \) were absent. To derive this expression for \( \dot{\bar{W}} = -\bar{W} \), we used (20) together with the fact that \( Q(x(t)) + R_u(u(t)) \) is bounded as \( \epsilon_H(t) \) 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,

\[
\frac{d|\bar{W}(t)|^2}{dt} = -2\bar{W}(t)^T N_{\text{om}} \leq -2\alpha\lambda_{\text{min}}(\Lambda)|\bar{W}|^2
\]

(25)

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

**Remark 1:** Typical adaptive optimal control algorithms [9] do not have the extra past-data terms \( \sum_{i=1}^k \frac{\omega(t_i)^T \omega(t_i)}{(\omega(t_i)^T \omega(t_i) + 1)^2} \) in the error dynamics (22) and thus need a persistence of excitation condition on \( \left( \frac{\omega(t)^T \omega(t)}{(\omega(t)^T \omega(t) + 1)^2} \right) \) (typically of the form \( \beta_1 I \leq \int_{t}^{t+T} \left( \frac{\omega(t)^T \omega(t)}{(\omega(t)^T \omega(t) + 1)^2} \right) \leq \beta_2 I \) with constants \( \beta_1, \beta_2, T \in \mathbb{R}^+ \)) that holds for every \( t \) from \( t = 0 \) to \( t = \infty \). This condition cannot be verified during learning. In Theorem 2, the 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 \( \Lambda \) being positive definite.

In practice, as one collects each additional vector \( \omega(t_i) \), one adds a new term to the matrix \( \Lambda \) and one can stop recording points as soon as this matrix becomes full-rank (i.e. \( t_i \) 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 (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

\[
\mathcal{L} = \frac{1}{2\alpha} \bar{W}(t)^T \bar{W}(t).
\]

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

\[
\dot{\mathcal{L}} = -\bar{W}(t)^T \left( \frac{\omega(t)^T \omega(t)}{(\omega(t)^T \omega(t) + 1)^2} + \Lambda \right) \bar{W}(t) + \bar{W}(t)^T P_{\text{err}},
\]

(27)

which is negative definite as long as,

\[
|\bar{W}| > \frac{|P_{\text{err}}|}{\lambda_{\text{min}}(\Lambda)}.
\]

Equation (25) follows from this and the fact that \( \frac{\omega(t)^T \omega(t)}{(\omega(t)^T \omega(t) + 1)^2} > 0, \forall t \). Since \( \{\omega(t_1), \ldots, \omega(t_k)\} \) has \( N \) linearly independent vectors, the matrix \( \Lambda \) 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 [42] to approximate both \( V^* \) and its gradient \( \nabla V^* \), but instead we independently adjust two sets of weights: the critic weights introduced in (17) to approximate \( V^* \) and the actor weights introduced below to approximate \( \nabla V^* \) 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,

\[ u^*(x) = W^*_u \phi_u(x) + \epsilon_u(x), \forall x \]  \hspace{1cm} (28)

where \( W^*_u \in \mathbb{R}^{N_2 \times m} \) 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 \( \epsilon_u \) is the actor approximation error. As before, the NN activation functions must define a complete independent basis set so that \( u^*(x) \) can be uniformly approximated on \( \Omega \), as expressed by the following assumption.

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

Since the ideal weights \( W^*_u \) are not known, we introduce actor estimate weights \( \hat{W}_u \in \mathbb{R}^{N_1 \times m} \) to approximate the optimal control in (28) by the following estimate:

\[ \hat{u}(x) = \hat{W}_u^T \phi_u(x), \forall x. \]  \hspace{1cm} (29)

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

\[ E_u(t) = \frac{1}{2} \text{trace}\{e_u^T(t)e_u(t)\}, \forall t, \]  \hspace{1cm} (30)

where

\[ e_u := \hat{W}_u^T \phi_u + \theta \left( \frac{1}{2} R^{-1} g^T(x) \nabla \phi^T \hat{W} \right) \in \mathbb{R}^m. \]

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

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

\[ \dot{\hat{W}}_u = -\alpha_u \frac{\partial E_u}{\partial \hat{W}_u} = -\alpha_u \phi_u e_u \]

\[ = -\alpha_u \phi_u \left( \hat{W}_u^T \phi_u + \theta \left( \frac{1}{2} R^{-1} g^T(x) \nabla \phi^T \hat{W} \right) \right)^T, \]  \hspace{1cm} (31)

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

\[ \tilde{W}_u := W^*_u - \hat{W}_u \in \mathbb{R}^{N_1 \times m}, \]  \hspace{1cm} (32)

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

\[ \tilde{W}_u = -\alpha_u \phi_u \theta \left( \frac{1}{2} R^{-1} g^T(x) \nabla \phi^T \hat{w}^* \right)^T + \alpha_u \phi_u \theta \left( \frac{1}{2} R^{-1} g^T(x) \nabla \phi^T \hat{w} \right)^T - \alpha_u \phi_u \theta. \]  \hspace{1cm} (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 \( \delta(\cdot) \), this term is always bounded and will be treated appropriately in the stability analysis that follows.

A pseudocode that describes the proposed adaptive-optimal control algorithm has the following form.

**Algorithm 2:** Adaptive-Optimal Control Algorithm with Relaxed PE

1. **procedure**
2. **Start** with initial state \( x(0) \) and random initial weights \( \hat{W}_u(0), \hat{W}(0) \)
3. **Propagate** \( t, x(t) \)
4. **Propagate** \( \hat{W}_u(t), \hat{W}(t) \) \( \to \hat{W}_u \) as in (31) and \( \hat{W} \) as in (20).
5. **Compute** \( \tilde{V}(x) = \hat{W}^T \phi(x) \) \( \to \) output of the Critic NN, and \( \hat{u}(x) - \hat{W}_u^T \phi_u(x) \) \( \to \) output of the Actor NN
6. **while** \( i \neq k \) **do** **Until** \( \{\omega(t_1), \omega(t_2), \ldots, \omega(t_k)\} \) has \( N \) linearly independent elements
7. **Select** arbitrary data points to be included in the history stack
8. \( i := i + 1 \)
9. **end while**
10. **end procedure**

**Remark 4:** The complexity of the above optimal-adaptive control algorithm is similar to other adaptive control algorithms [33] and for that reason, it increases with the number of states \( n \). Also, note 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 [35]. 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 \( \Omega \subseteq \mathbb{R}^n \), i.e. \( |g(x)| < 1/2, \forall x \in \Omega \).

To remove the effect of the NN approximation errors \( \epsilon, \epsilon_u \) (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 [43] and use:

\[ \tilde{u}(x) = \hat{W}_u^T \phi_u(x) + \eta, \forall x, \]  \hspace{1cm} (34)

where

\[ \eta := -B \frac{|x|^2}{(A + x^T x)}, \forall x, \]  \hspace{1cm} (35)

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

\[ B |x|^2 \geq \frac{A + x^T x}{(W_m \phi_{dm} + \epsilon_{dm})} \left\{ \frac{1}{2\alpha} \left( \frac{k + 1}{2} \epsilon_{Hm} \right)^2 + (\phi_{um} \bar{u} + \phi_{um} \epsilon_{um})^2 + \frac{\phi_{um}}{2} (W_m \phi_{dm} + \epsilon_{dm})^2 \right\} + 2(2 \phi_{um} \bar{u}^2 + (W_m \phi_{dm} + \epsilon_{dm})^2 + \epsilon_{um}^2). \]  \hspace{1cm} (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):

$$\dot{x} = f(x) + g(x)((W_u^* - \tilde{W}_u)^T \phi_u(x) + \eta),$$ (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 \(\{\omega(1), \omega(2), \ldots, \omega(k)\}\) has \(N\) linearly independent elements. Then, there exists a triple \((\Omega_x \times \Omega_W \times \Omega_{W_u}) \subseteq \Omega \times \Omega\) compact such that the solution \(\tilde{Z} := (x(t), \tilde{W}(t), W_u(t)) \in (\Omega_x \times \Omega_W \times \Omega_{W_u})\) exists globally and converges asymptotically to zero for all neural network weights \(\tilde{W}(0)\) inside \(\Omega_W\), \(\tilde{W}_u(0)\) inside \(\Omega_{W_u}\) and state \(x(0)\) inside \(\Omega_x\), provided that the following inequalities are satisfied,

$$\begin{align*}
\alpha &> \sqrt{\frac{1}{8\lambda_{\min}(\sum_{i=1}^{k} \frac{\omega(t_i)^2}{\omega(t_i) + 1})}} \\
\phi_{um} &> \frac{1 + \sqrt{65}}{8}.
\end{align*}$$ (38) (39)

When the set \(\Omega\) 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 5:** For the inequality (38) to hold, one needs to pick the tuning gain \(\alpha\) for the critic NN sufficiently large. But as noted in adaptive control [33], large adaptive gains can cause high frequency oscillations in the control signal and reduced tolerance to time delays that 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 \(\phi_{um}\) is simply an upper bound that appears in Assumptions 2, one can have it as large as needed. However, one must keep in mind that a large value for \(\phi_{um}\), requires an appropriate value for the function \(B\) in the robustness term in (35), (38)

**Remark 6:** By denoting as \(\tilde{Z} := [x^T \tilde{W}^T \tilde{W}_u^T]^T\) \(\tilde{Z}\) from the conclusion of Theorem 3 we have that \(|\tilde{Z}|\to 0\) which implies \(|x|\to 0\), it is straightforward that as \(t\to\infty\) then from (34) we have (29).

**Remark 7:** 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 \(V^*\) is norm coercive (i.e., \(V^*(x) \to 0 \Rightarrow x \to 0\)), as this suffices to guarantee that the Lyapunov function \(V\) 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

$$V := V^* + \tilde{W}^T \tilde{W} + \frac{1}{2\alpha_u} \text{trace}\{\tilde{W}_u^T \tilde{W}_u\},$$ (40)

where \(V^*\) is the optimal value function in (5) that is the positive definite and smooth solution of (8) (see Theorem 1), and \(V_c := W^T W\) is the Lyapunov function considered in

**Theorem 2:** Since \(\mathcal{V}\) is positive definite, there exist class-\(\mathcal{K}\) functions \(\gamma_1(\cdot)\) and \(\gamma_2(\cdot)\), then,

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

for all \(\tilde{Z} := [x^T \tilde{W}^T \tilde{W}_u^T]^T \in B_r\), 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,

$$\dot{\mathcal{V}} = \nabla V^* T (f(x) - g(x) \tilde{W}_u^T \phi_u + + g(x)(u^* - \epsilon_u) - g(x)B |x|^2 \frac{1}{A + x^T x})$$

$$\begin{align*}
&- \dot{\mathcal{V}}_c - \frac{\omega(t)^2}{\omega(t) + 1} + \frac{\omega(t)^2}{\omega(t) + 1} T + \\
&\frac{\omega(t)^2}{\omega(t) + 1} T + \frac{\omega(t)^2}{\omega(t) + 1} T + \\
&+ \text{trace}\{\tilde{W}_u^T (- \phi_u T \tilde{W}_u - \phi_u \theta \frac{1}{2} R^{-1} g(x) \nabla \phi_T \tilde{W})^T
- \phi_u \theta \frac{1}{2} R^{-1} g(x) \nabla \epsilon)^T - \phi_u \epsilon_u), \ t \geq 0,
\end{align*}$$

$$= T_1 + T_2 + T_3.$$ (41)

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

Using the HJB equation,

$$\nabla V^* T f(x) = -\nabla V^* T g(x) u^* - R_u(u^*) - Q(x), \ \forall x$$

in (44) yields,

$$T_3 = -R_u(u^*) - Q(x) - \nabla V^* T g(x) \tilde{W}_u^T \phi_u - \nabla V^* T g(x) \epsilon_u$$

$$\begin{align*}
- \nabla V^* T g(x) B |x|^2 \frac{1}{A + x^T x} &\leq -R_u(u^*) - Q(x) \\
- (W_m \phi_{dm} + \epsilon_{dm}) \left(\frac{1}{2} \phi_{um} |\tilde{W}_u| + \\
+ (\epsilon_{um} + \frac{1}{2} B |x|^2 \frac{1}{A + x^T x}) \right) (45)
\end{align*}$$

since \(A + x^T x > 0\). The term \(T_3\) can be further upper bounded as,

$$T_3 \leq -R_u(u^*) - Q(x) + \frac{\phi_{um}}{4} ((W_m \phi_{dm} + \epsilon_{dm}))^2$$

$$+ \frac{\phi_{um}}{4} |\tilde{W}_u|^2 + \frac{1}{2} (W_m \phi_{dm} +$$

$$\epsilon_{dm})^2 + \frac{1}{2} \epsilon_{um} - (W_m \phi_{dm} + \epsilon_{dm}) \frac{1}{2} B |x|^2 \frac{1}{A + x^T x}. (46)$$

Finally after taking into account the bound of \(B |x|^2\) from (36) we can upper bound (41) as,
T_1 := - \frac{\partial V_c}{\partial W} T \left( \left( \omega(t) \omega(t) \right)^T + \sum_{i=1}^{k} \omega(t_i) \omega(t_i) T \right) \hat{W} + \frac{\partial V_c}{\partial W} \left( \frac{\omega(t)}{(\omega(t) \omega(t) + 1)^2} \epsilon_H(t) + \sum_{i=1}^{k} \omega(t_i) \right) 
\leq -2\alpha \lambda_m (\sum_{i=1}^{k} \frac{\omega(t_i) \omega(t_i) T}{(\omega(t_i) \omega(t_i) + 1)^2} |\hat{W}|^2 + \frac{1}{2\alpha} |\hat{W}| \left( \frac{k + 1}{2} \epsilon_H \right) 
\leq -2\alpha \lambda_m (\sum_{i=1}^{k} \frac{\omega(t_i) \omega(t_i) T}{(\omega(t_i) \omega(t_i) + 1)^2} |\hat{W}|^2 + \frac{1}{4\alpha} |\hat{W}|^2 + \frac{1}{4\alpha} \left( \frac{k + 1}{2} \epsilon_H \right)^2 \right) \quad (42)

T_2 := \text{trace} \{ \hat{W} \hat{V} - \phi_u \phi_u^T \hat{W} \} \left( \frac{1}{2} R^{-1} g^T(x) \nabla \phi \hat{V} \right)^T 
+ \phi_u \left( \frac{1}{2} R^{-1} g^T(x) \nabla \phi \hat{W} \right)^T 
\leq -\phi_u^2 |\hat{W}|^2 + 2\phi_u \mu |\hat{W}| + (\phi_u \mu + \phi_u \mu_u) |\hat{W}| 
\leq -\phi_u^2 |\hat{W}|^2 + (2\phi_u \mu)^2 + \frac{(\phi_u \mu + \phi_u \mu_u)^2}{2} + |\hat{W}|^2. \quad (43)

T_3 := \nabla V^* T \left( f(x) - g(x) \hat{W} \phi_u + g(x)(u^* - \epsilon_u) - g(x) B |x|^2 \frac{1}{A + x^T x} \right) \quad (44)

\dot{V} \leq -2\alpha \lambda_m (\sum_{i=1}^{k} \frac{\omega(t_i) \omega(t_i) T}{(\omega(t_i) \omega(t_i) + 1)^2} - \frac{1}{4\alpha}) |\hat{W}|^2 
- \left( \phi_u^2 - \frac{\phi_u \mu}{4} - 1 \right) |\hat{W}|^2 
- R_u(u^*) - Q(x), \quad t \geq 0. \quad (47)

Then, by taking into account the inequalities (38) and (39) one has, \( V \leq 0, \quad t \geq 0 \). From Barbalat’s lemma [44] it follows that as \( t \to \infty \), then \( \dot{Z} \to 0 \).

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

\[ M := \{ x \in \mathbb{R}^n | V(t) \leq m \} \subset \mathbb{R}^n \]

where \( m \) is chosen as the largest constant so that \( M \subseteq \Omega \). Since by assumption \( x_0 \in \Omega \), and \( \Omega_x \subset \Omega \) then we can conclude that \( x_0 \subset \Omega \). While \( x(t) \) remains inside \( \Omega \), we have seen that \( \dot{V} \leq 0 \) 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.

V. SIMULATIONS

This section presents two simulation examples to illustrate the effectiveness of the proposed optimal adaptive control algorithm.

A. Van-der Pol Oscillator

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

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

which has an uncontrolled unstable limit cycle and a stable uncontrollable equilibrium point at the origin. It is shown in [45] that the non-saturated optimal control input, with a criterion \( V = \int_0^T (|x|^2 + |u|^2) dt \), is given by \( \hat{u} = -x_1 - x_2 \) and that the corresponding optimal value function is given by \( V^* = x_1^2 + x_2^2 \).

We now consider the optimal control of (48) with the input saturated so that \( |u| \leq \bar{u} = 0.1 \) and the cost defined by (2), (3), and (4) with \( Q(x) := |x|^2 \) 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 \( \phi(x) = [x_1^2 x_2 x_1^2]^T \) and \( \hat{\phi}(x) = \nabla \phi(x) \), and the tuning gains were set to \( \alpha = 10, \alpha_u = 2 \). Thus, the critic parameters converged to \( \hat{W} = [2.4953 0.9991 2.2225] \).

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.

B. Power Plant System

Consider the power system shown in Figure 3, 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

\[ \left[ \Delta \hat{\alpha} \Delta \hat{\phi} \right] \Delta \hat{P}_m \] \[ = \left[ \begin{array}{ccc} -\frac{1}{T_\alpha} & 0 & \frac{1}{T_\phi} \\ \frac{1}{T_\phi} & 0 & -\frac{1}{T_\phi} \end{array} \right] \left[ \Delta \hat{\alpha} \Delta \hat{\phi} \right] \] \[ + \left[ \begin{array}{c} -\frac{1}{T_\phi} \\ 0 \end{array} \right] \Delta \hat{P}_m. \quad (49) \]

where \( \Delta f_G \) is the incremental frequency deviation, \( \Delta P_m \) is the incremental change in the generator output, \( \Delta \hat{\alpha} \) is
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 $|\Delta P_c| \leq \bar{u} = 0.02$ and the cost defined by (2), (3), and (4) with $Q(x) := \Delta \bar{\alpha}^2 + \Delta P_m^2 + \Delta f_G^2$ and $R = 0.5$. The initial NN weights were randomly initialized in $[0, 1]$, tuning gains set to $\alpha = 10$, $\alpha_u = 2$, and the activation functions were chosen to be quadratics of the form $\phi(x) = [x_1^2 x_1 x_2 x_1 x_3 x_2^2 x_2 x_3 x_3^2]^T$, $\phi_u(x) := \nabla \phi(x)$. The critic neural network weights converged to $\hat{\mathbf{W}} = [0.0583 \ 0.0476 \ 0.0549 \ 0.1123 \ 0.1447 \ 0.3489]$. Figure 4 shows the state evolution, Figure 5 the incremental speed change in position deviation (control input), and Figure 6 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.

The incremental change in governor value position, and the control input $\Delta P_c$ of the system is the incremental speed change in position deviation [46]. 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 $K_t = 1[s]$. 
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.

REFERENCES
