An Online Single-Network Adaptive Algorithm for Continuous-Time Nonlinear Optimal Control

Jae Young Lee¹, Jin Bae Park¹*, Yoon Ho Choi², and Keun Uk Lee¹

¹Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
(Tel : +82-2-2123-2773; E-mail: {jyoungle, lkucj, jbpark}@yonsei.ac.kr)

²Department of Electronic Engineering, Kyonggi University, Suwon, Kyonggi-Do, Korea
(Tel : +82-31-249-9801; E-mail: yhchoi@kyonggi.ac.kr)

Abstract: In this paper, we propose an online adaptive neural-algorithm to solve the CT nonlinear optimal control problems. Compared to the existing methods, which adopt the architecture with two neural networks (NNs) for actor-critic implementations, only one NN for critic is used to implement the algorithm, simplifying the structure of the computation model. Moreover, we also provide a generalized learning rule for updating the NN weights, which covers the existing critic update rules as special cases. The theoretical and numerical results are given under the required persistent excitation condition to verify and analyze stability and performance of the proposed method.

Keywords: Adaptive control; optimal control; actor-critic; approximate dynamic programming; nonlinear control

1. INTRODUCTION

In the several decades, optimal control theories have been developed for the design of modern control systems with successful applications [1, 2]. However, intractability of the associated Hamilton-Jacobi-Bellman (HJB) equation has raised difficulties in the analytical design of nonlinear optimal control. To alleviate these difficulties, several numerical methods have been provided by control and computer scientists [3–13].

In the fields of reinforcement learning and approximate dynamic programming, forward-in-time computation algorithms called actor-critic methods or adaptive critics have been developed during last two decades [6–13]. These methods solve the associated HJB equation in forward time to obtain the approximate neural network (NN) optimal control solution. After the pioneering works of Werbos [8] and Prokhorov et al. [9] for discrete-time general nonlinear systems, many researchers have provided their own actor-critic algorithms to solve nonlinear optimal control problems associated with input-affine nonlinear systems. Although the input-affine nature enables to obtain the control policy directly from the value function approximated by the critic NN, most of those algorithms use the additional actor NN to approximate the control policy.

From the input-affine nature, Padhi et al. [10] proposed a single-network adaptive critics (SNAC) NN architecture for discrete-time input-affine nonlinear systems. The algorithm has just one NN called critic to approximate the value function, and does not need additional actor NN, reducing the computational cost; the convergence of the algorithm was also given in [10]. For continuous-time case, the adaptive actor-critic methods utilizing two NN structures are proposed in [11–13]. While adaptive critics method for non-affine system employs the actor NN to approximate the control policy [11], the actor NNs in the input-affine actor-critic algorithms are introduced to guarantee closed-loop stability [12, 13].

Motivated by the works [12] and [13], this paper proposes an online adaptive neural-algorithm that utilizes the critic NN only, to solve the CT nonlinear optimal control problems; Compared with the other actor-critic methods in CT, the proposed method does not necessitate the additional actor NN structure, simplifying the structure of the computation model as in SNAC proposed in [10]. Moreover, the proposed learning rule for updating the NN weights is generalized from and as special cases, covers the critic update rules given in [12] and [13]. The theoretical and numerical results are given under the required persistent excitation condition, to verify and analyze stability and performance of the proposed method.

2. PROBLEM FORMULATION

Consider the following input-affine nonlinear system

\[
\dot{x} = f(x) + g(x)u, \quad x(0) = x_0
\]

with the infinite-horizon performance index

\[
J(x_0, u(\cdot)) = \int_0^{\infty} x^T Q x + u^T R u \, d\tau
\]

where \(x \in \mathbb{R}^n\) and \(u \in \mathbb{R}^m\) are the state variable and control input of the system (1); \(f : \mathcal{D} \to \mathbb{R}^n\) and \(g : \mathcal{D} \to \mathbb{R}^{n \times m}\) are nonlinear functions that are Lipschitz continuous on a given domain \(\mathcal{D} \subseteq \mathbb{R}^n\) that contains a neighborhood of the origin; \(Q \in \mathbb{R}^{n \times n}\) and \(R \in \mathbb{R}^{m \times m}\) are positive definite state and control-input penalty matrices, respectively. For a well-posed problem, we assume \(f(x) + g(x)u(x)\) is Lipschitz continuous on \(\mathcal{D}\), \(f(0) = 0\), and there is an admissible state-feedback policy \(u_a(x)\)
that asymptotically stabilizes the system (1) and guarantees the value function $V^*(x_0) := J(x_0, u_0)|_{u_0 = u_a}$ finite for all $x_0 \in D$.

The objective of the target algorithm is to find the optimal policy $u^*$ and the corresponding optimal value function $V^*(x)$ in a stable manner, where $V^*(x)$ is defined as

$$V^*(x) := J(x_0, u^*)|_{u = u^*},$$

(3)

and satisfies $0 \leq V^*(x) \leq V^{u_a}(x)$ for all $x \in D$ and all admissible policy $u_a(x)$. Define the Hamiltonian $H$ as

$$H(x, u, p) := p^T (f(x) + g(x)u_a) + x^T Qx + u_a^T Ru_a.$$

Then, assuming $V^{u_a}$ is $C^1$, an admissible policy $u_a(x)$ satisfies the Hamiltonian equation

$$H(x, u_a, \nabla V^{u_a}) = 0, \quad \forall x \in D.\quad (4)$$

Moreover, by Pontryagin’s minimum principle, the optimal control problem (1)–(2) can be solved by minimizing the Hamiltonian $H(x, u_a, \nabla V^{u_a})$ among all admissible policy $u_a(x)$. Differentiating $H(x, u, \nabla V^{u_a})$ with respect to $u_a$ and equating zero, we obtain the optimal policy $u^*$ given by

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

Substituting this into (4) and rearranging the resultant equation yield the HJB equation of the form:

$$0 = x^T Qx + \nabla T V^* f(x) - \frac{1}{4} \nabla T V^* g(x) R^{-1} g^T(x) \nabla V^*.$$  

(5)

In this paper, we assume there exists $V^* \in C^1$ that satisfies (5) for all $x \in D$, which is the necessary and sufficient condition for optimality.

3. MAIN RESULTS

In this section, we propose an online adaptive learning method that uses only one NN called critic to find the approximate solution of $V^*$ and $u^*$. As a first step, note that $V^* \in C^1$ can be written as

$$V^*(x) = (W^*)^T \phi(x) + \epsilon(x)$$

(6)

where $W^* \in \mathbb{R}^N$ is the NN optimal weight vector, $\phi : D \rightarrow \mathbb{R}^N$ is the critic NN activation function, and $\epsilon(x) \in \mathbb{R}$ is the reconstruction error. Since the optimal weight vector $W^*$ is not known a priori, we approximate $W^*$ by its estimate $\hat{W}$, so the NN structure of the proposed method is given by

$$\hat{V}(x) = \hat{W}^T \phi(x),$$

(7)

where $\hat{V}(x)$ is the NN output, an estimate of the value function. From this, we define the approximated control $\hat{u}$ as

$$\hat{u} := -\frac{1}{2} R^{-1} \nabla T \hat{V}(x).$$

(8)

Substituting (7) and (8) into the system (1) under $u = \hat{u}$, we have the closed-loop system expressed as

$$\dot{x} = f(x) - \frac{1}{2} g(x) R^{-1} g^T(x) \nabla T \phi(x) \hat{W}.$$  

(9)

The objective of the proposed method is to make $\hat{u} \approx u^*$ by updating the NN weight vector $\hat{W}$. Note that the reconstruction error $\epsilon(x)$ in (6) is bounded in a given compact set by universal approximation theorem if $N$ is sufficiently large. Substituting $u_a = \hat{u}$ and $V^{u_a} = \hat{V}$ into (4), we obtain

$$\sigma T(x) \hat{W} + x^T Qx + \hat{u}^T R \hat{u} = \delta_{HJB},$$

(10)

where $\sigma(x) = \nabla T \phi(x)(f(x) + g(x)\hat{u})$, and $\delta_{HJB}$ denotes the Bellman error indicating the error of the Hamiltonian equation (4) due to the approximation of the policy and value function. In the proposed method, $\hat{W}$ will be updated by the following update rule.

$$\hat{W} = -\eta \Gamma \frac{\sigma}{(1 + \gamma \sigma^T \Gamma \sigma)^k} \cdot \delta_{HJB}$$

$$= -\eta \Gamma \frac{\sigma}{(1 + \gamma \sigma^T \Gamma \sigma)^k} \left[ \sigma T \hat{W} + x^T Qx + \hat{u}^T R \hat{u} \right],$$

(11)

where $\eta > 0$ is the learning rate; $k \in \mathbb{Z}_+$ and $\gamma > 0$ are constants; $\Gamma(t) \in \mathbb{R}^{N \times N}$ is a bounded positive definite gain matrix which is either constant or scheduled by the learning agent. When $k = 2$, $\Gamma = I$, and $\gamma = 1$, (11) becomes exactly same to the normalized gradient descent method given in [12], which minimizes the squared Bellman error $E_G(\delta_{HJB}) = \frac{1}{2} \delta_{HJB}^2$. On the other hand, if $k = 1$, and $\Gamma$ is generated by

$$\hat{\Gamma} = -\eta \Gamma \frac{\sigma T}{1 + \gamma \sigma^T \Gamma \sigma} \Gamma,$$

(12)

with an initial condition $\Gamma(0) = \alpha I > 0$ and covariance resetting $\Gamma(t^+) = \Gamma(0)$ (where $t^+$ is the resetting time at which the minimum eigenvalue of $\Gamma$ becomes less than $\alpha$), then (11) and (12) are the normalized least squares in the critic update rule given in [13]. In this case, the update rule minimizes the integral squared Bellman error $E(\delta_{HJB})$:

$$E(\delta_{HJB}(x(t))) := \int_0^t \delta_{HJB}^2(x(\tau)) \, d\tau.$$

Remark 1: The main advantage of the proposed adaptive algorithm with respect to the actor-critic methods given in [12, 13] is that ours does not require the
second actor NN, so the computational cost is reduced. Moreover, the update rule (11) is generalized and contains both critic update rules in [12] and [13] as special cases.

To correctly learn the optimal parameters \( W^* \) while maintaining stability, the following persistently exciting condition is necessary for \( \psi(t) \in \mathbb{R}^n \) defined by
\[
\psi(t) := \sigma / (1 + \gamma \sigma^T \Gamma \sigma)^k / 2.
\]

**Assumption 1**: \( \psi(t) \) is persistently exciting. That is, there are some constants \( g, \alpha > 0 \), and \( T > 0 \) such that for all \( t \geq 0 \),
\[
\alpha I \leq \int_{t}^{t+T} \psi(\tau)\psi^T(\tau) \, d\tau \leq \alpha I. \tag{13}
\]

Let \( \hat{W} \) be the NN weight error vector defined by \( \hat{W} := \bar{W} - W^* \). Then, the error dynamics of \( W \) can be derived from (10) and (11) as follows:
\[
\dot{\hat{W}} = -\eta \Gamma \psi \hat{W} - \eta \Gamma \sigma (1 + \gamma \sigma^T \Gamma \sigma)^k \Delta(x, \hat{W}), \tag{14}
\]
where \( \Delta(x, \hat{W}) := \hat{W}^T \phi g R^{-1} g^T \hat{W} + \delta_{H, IB} \) denotes the perturbation term. Note that under the PE condition (13), the nominal system
\[
\dot{\tilde{W}} = -\eta \Gamma \psi \tilde{W}
\]
is exponentially stable [14, Theorem 2.5.1]. Based on this property, the following statement regarding closed-loop stability can be established.

**Theorem 1**: Assume that \( g(x), \phi(x), \) and \( \varepsilon(x) \) are bounded. Then, under Assumption 1, there exists a constant \( c > 0 \) such that for all \( (x(0), \hat{W}(0)) \in \mathcal{B}_{c} \), where \( \mathcal{B}_{c} \) is defined as
\[
\mathcal{B}_{c} := \{ (x, W) \in \mathbb{R}^{n \times N} : \|x\|^2 + \|W\|^2 \leq c \},
\]
the combined closed-loop system (9) and (14) is uniformly ultimately bounded.

**Proof**: The proof is omitted due to space limitation. 

4. SIMULATIONS

To verify the performance, we consider the following linear system:
\[
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2
\end{bmatrix} =
\begin{bmatrix}
-1 & -2 \\
1 & -4
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} +
\begin{bmatrix}
1 \\
-3
\end{bmatrix} u \equiv Ax + Bu \tag{15}
\]
with identity penalty matrices \( Q = I \) and \( R = 1 \).

This system is also simulated in [11] to verify the performance of the adaptive critics method. In this linear case, the value function is given in a quadratic form
\[
V^*(x) = x^T P x,
\]
so the HJB equation (5) becomes the algebraic Riccati equation of the form
\[
A^T P + PA - PBR^{-1}B^T P + Q = 0.
\]
In the simulation, the quadratic value function is parameterized as
\[
V^*(x) = (W^*)^T \phi(x) \quad \text{with} \quad \phi(x) = [x_1^2, x_1 x_2, x_2^2],
\]
where \( W^* \) consists of the elements of \( P \), and given by \( W^* = [0.0199, -0.1162, 0.1292]^T \).

The parameters \( \Gamma, \gamma, \) and \( k \) in the learning rule (11) are set to \( \Gamma = I, \gamma = 1, \) and \( k = 2 \) to consider the normalized gradient descent case [12]. The learning rate and the initial condition in the simulation are given by \( \eta = 10 \) and \( (x(0), \hat{W}(0)) = (0, 0, 0, 0), \) respectively. To excite \( \psi(t) \) for satisfying PE (13), the exploration signal \( n(t) = 2(\sin^2 t \cos t + \sin^2 2t \cos 0.1 t + \sin^2 t - 1.2 t \cos 0.5 t + \sin^2 t) \) is applied to the system (1) through the control input channel. This exploration signal \( n(t) \) is vanished after \( \hat{W}(t) \) converges (at \( t = 45 \) [s] in the simulation).

Fig. 1 demonstrates the variations of the NN weights \( \hat{W} \), which definitely converge as shown in the figure. On the contrary, the trajectories of \( x \) shown in Fig. 2 are rather oscillatory during the learning phase, but converges to zero after eliminating the exploration \( n(t) \). The final weights \( \hat{W}(t_f) \) at \( t_f = 50 \) [s] is
\[
\hat{W}(t_f) = [0.0198, -0.1158, 0.1293]^T,
\]
which is almost same to \( W^* \) and shows the effectiveness of the proposed single-NN adaptive algorithm.

5. CONCLUSIONS

In this paper, we proposed an online adaptive neural-algorithm to solve the CT nonlinear optimal control problems. In the proposed method, one NN, called critic, was
utilized to approximate the optimal value function by online learning, and additional actor NN was not employed as opposed to the learning methods given in [12, 13]. The proposed generalized learning rule for updating the NN weights contains the existing critic update rules as special cases [12, 13] and guarantees the uniform ultimate boundedness of the closed-loop learning system. Numerical simulations were performed to support the theoretical results and verify the effectiveness of the proposed method.

6. ACKNOWLEDGEMENTS

This work has been supported by National Research Foundation of Korea Grant funded by the Korean Government (No. NRF-2013R1A1A2012851).

REFERENCES