Neural Networks 32 (2012) 229–235

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2012 Special Issue

A boundedness result for the direct heuristic dynamic programming

Feng Liu [a], Jian Sun [a], Jennie Si [b,*], Wentao Guo [a], Shengwei Mei [a]

[a] Department of Electrical Engineering, Tsinghua University, Beijing, 100084, PR China
[b] Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287, USA

ARTICLE INFO

Keywords:
Approximate dynamic programming (ADP)
Direct heuristic dynamic programming (direct HDP)
Lyapunov stability
Uniformly ultimately boundedness (UUB)

ABSTRACT

Approximate/adaptive dynamic programming (ADP) has been studied extensively in recent years for its potential scalability to solve large state and control space problems, including those involving continuous states and continuous controls. The applicability of ADP algorithms, especially the adaptive critic designs has been demonstrated in several case studies. Direct heuristic dynamic programming (direct HDP) is one of the ADP algorithms inspired by the adaptive critic designs. It has been shown applicable to industrial scale, realistic and complex control problems. In this paper, we provide a uniformly ultimately boundedness (UUB) result for the direct HDP learning controller under mild and intuitive conditions. By using a Lyapunov approach we show that the estimation errors of the learning parameters or the weights in the action and critic networks remain UUB. This result provides a useful controller convergence guarantee for the first time for the direct HDP design.

1. Introduction

In the past decades, the adaptive/approximate dynamic programming (ADP) has attracted intensive attention from researchers attempting to obtain approximate solutions to the Hamilton–Jacobi–Bellman (HJB) equation (Balakrishnan, Jie & Lewis, 2008; Bellman & Dreyfus, 1962; Si, Barto, & Powell, 2004). Primary results on ADP designs with convergence guarantees usually were obtained for discrete-time finite state and finite control problems under a Markov decision process (MDP) framework. The well known Q-learning method by Watkins and Dayan (1992) is one such case with its convergence analysis performed based on an MDP formulation with finite state and finite control constraint imposed. But the scalability of the algorithm is questionable. The temporal difference (TD) method introduced a new way of approximating the value function by prediction based on temporal differences. Since the optimal control is still obtained by enumeration in most applications as analytical solutions are generally not attainable, TD methods do not completely overcome the challenge of scalability, despite its success in addressing the issue of approximating the value function. The actor critic techniques (Barto, Sutton, & Anderson, 1983) as well as adaptive critics (Werbos, 1977) opened new windows by introducing approximating functions into the iterative process of updating the control policy and value function estimates. They become popular and have shown great promise to address the curse of dimensionality.

The heuristic dynamic programming (HDP) (Werbos, 1989), one of the adaptive critic design techniques, creates a whole new way of computing both the value function and the optimal control by approximations. With this idea, the curse of dimensionality is finally, and in a principled way, possible to be addressed. To actually compute the optimal solutions efficiently and reliably, Werbos proposed additional adaptive critic structures: the dual heuristic programming (DHP) and the globalized DHP (Werbos, 1992). As the structures become more complex, the design aims at accommodating more complex terms in the computation of the underlying optimal control.

Several applications have been demonstrated by using adaptive critic methods. In Prokhorov, Santiago, and Wunsch (1995), the series of adaptive critic design approaches were examined and tested on the problem of a simulated, simple aircraft landing control. The results show that tighter control can be obtained with more complex adaptive critic techniques. Results of adaptive critic optimal control for a missile guidance problem also illustrated the potential of adaptive critic methods for continuous state and continuous control problems. Among others, additional illustrative examples of adaptive critic designs including power system excitation and stability control problems (Ernst, Glavic, & Wehenkel, 2004; Venayagamoorthy, Harley, & Wunsch, 2003). Those simulation results show great promise of adaptive critic methods to scalable applications on either continuous or discrete state control problems. The natural concern, however, is how to develop stable learning control algorithms with convergence or performance guarantees.

The direct HDP was developed in Si and Wang (2001). It takes advantage of the potential scalability of the adaptive critic designs

* Corresponding author.
E-mail address: si@asu.edu (J. Si).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.02.005

and the intuitiveness of Q-learning. It is an online learning scheme that simultaneously updates the value function approximation and the action function approximation. It has been applied to various realistic and complex control problems that involve continuous states and controls. In Enns and Si (2002, 2003a, 2003b), the direct HDP controller was designed to stabilize, maneuver, and reconfigure after fault an Apache helicopter using a full, industrial scale simulator under difficult and realistic flying conditions. In another report (Lu, Si, & Xie, 2008), the direct HDP was used to provide coordinated control for damping low frequency oscillations in a realistic large-scale power system. Most recently, a PID structure embedded in the direct HDP controller design was proposed in Sun, Liu, and Si (2011), and it was demonstrated that this design may potentially improve the convergence and robustness properties of online controllers based on the direct HDP.

Previous case studies are encouraging in the sense that adaptive critic designs and the direct HDP design have shown their promising potential to scale up for large-scale, realistic control problems. Despite its relative ease and reliability in implementation, adaptive critic designs, direct HDP, or other co-adaptation structures that involve simultaneous learning of both action and critic/value functions, there is no readily-available theoretical guarantee of convergence in learning parameters under general conditions.

In this paper, a Lyapunov stability approach is used for the analysis of the direct HDP learning control algorithm. It is shown that the estimation errors of the action and critic network weight parameters retain a uniformly ultimate boundedness (UUB) property under mild conditions. Since the result provides an explicit condition for the learning rates, it can be readily used in the training of direct HDP controllers as described in Si and Wang (2001).

The rest of this paper is organized as follows: Section 2 provides a backdrop of related work; Section 3 formulates the direct HDP in preparation for the analysis conducted next; the main theoretical results are presented in Section 4; Section 5 provides a simple example for illustration; finally, some concluding remarks are given in Section 6.

2. Related works

The convergence analysis of the actor critic co-adaptation was initially attempted under a linear problem formulation. In Bradtke, Ydestie, and Barto (1994), presented a proof for a linear system with a linear quadratic regulator using a special form of the Q-learning. It was further extended to using HDP and dual HDP with a similar linear control structure in Landelius (1997).

For continuous state and action spaces, convergence results are more challenging when using nonlinear function approximators. Lewis et al. made significant contributions in this regard. In Abu-Khalaf and Lewis (2005, 2008), Abu-Khalaf, Lewis, and Huang (2006), Vamvoudakis and Lewis (2009) and Vrabie and Lewis (2009), several analytical frameworks have been developed for ADP control designs under *special* nonlinear system formulations. Most of their results were based on a class of affine nonlinear control systems described by $\dot{x} = f(x) + g(x)u$. The ADP-based controllers were then developed based on such special system structure. In their work, neural networks were utilized to construct the action and the critic networks. To achieve online implementation, these two networks were simultaneously adapted. It was revealed that the convergence of the critic to the actual optimal value could be guaranteed if meeting the persistence of excitation condition. Moreover, the requirement to the explicit knowledge of the system can be avoided by using neural networks as universal approximators. This indicates the possibility of extending their special model-based results to the more general, model-free cases (Vamvoudakis & Lewis, 2009; Vrabie & Lewis, 2009).

However, the learning controllers considered in those results are limited to the imposed system structure. Specifically, (1) the controller designs require knowledge of $g(x)$ for deriving the optimal policy $u(x)$. In the suggested implementation, the action neural network is directly determined from the critic neural network by taking the gradient of the activation functions. As a result, the optimal policy $u(x)$ cannot be estimated directly by the action network. Instead, it can only be computed by using the explicit expression of $g(x)$; (2) the instantaneous cost function is required to be of certain special forms, for example, a quadratic function in $x(t)$ and $u(t)$. Their proposed online implementation was based on this explicit instantaneous cost function. Hence, to derive gradient descent algorithms for computing the value function and the optimal policy, the derivatives of the instantaneous cost function should be easily obtainable. In Abu-Khalaf and Lewis (2005, 2008), Abu-Khalaf et al. (2006), Vamvoudakis and Lewis (2009) and Vrabie and Lewis (2009), $r(x(t), u(t)) = Q(x(t)) + u(t)^T R u(t)$ with positive $Q(x(t))$ and R was used as the instantaneous cost function. As a result, these may hinder the applications of the online ADP controllers.

Different from the policy iteration approaches mentioned above, the direct HDP combines the policy iteration and value iteration in one actor critic structure (Si & Wang, 2001). As a result, the online controller design does not require explicit expression or knowledge of either $f(x)$ or $g(x)$, or any explicit system identification procedure a priori. It is therefore, a model-free or direct approach. Moreover, in applications under general problem formulation conditions, the primary cost or the instantaneous cost function in ADP can be less informative than a quadratic function based on states and controls. For instance, in cases of board games, the reward function is delayed and the value is only binary. As a result, the direct HDP approach in principle can be applied to a wide class of problems, including both explicit cost/reward or delayed cost/reward.

Given the above observations on the essential differences between the direct HDP and the controller designs in Abu-Khalaf and Lewis (2005, 2008), Abu-Khalaf et al. (2006), Vamvoudakis and Lewis (2009) and Vrabie and Lewis (2009), the results and the methodologies proposed therein cannot be directly extended to the design and analysis of the direct HDP online controllers. In Yang, Si, Tsakallis, and Rodriguez (2009), presented the direct HDP design in a nonlinear tracking control setting with filtered tracking error, a Lyapunov stability approach was used for the stability analysis of the tracking system based on a Brunovsky system model. The sufficient condition for the convergence result, however, was not intuitive or difficult to interpret. Moreover, it is not easy to use this result in controller design for applications. This motivates us to develop a convergence guarantee result for the direct HDP learning controller under reasonable conditions.

3. The direct heuristic dynamic programming

For the ease of discussion, we provide a brief overview of the direct HDP proposed in Si and Wang (2001), which also serves to define the convention used in this paper.

Consider a nonlinear dynamic system model of the following general form,

$$x(t+1) = f[x(t), u(t)] \tag{1}$$

where x denotes the system state, and u is the control variable.

As shown in Fig. 1, the direct HDP is structured so that its critic network or its value function approximation aims at solving the Bellman equation,

$$J^*(t) = \min_{u(t)} \left\{ r(t+1) + \alpha J^*(t+1) \right\}, \tag{2}$$