Acta Automatica Sinica
Vol. 39, No. 4 April, 2013

An Overview of Research on Adaptive Dynamic Programming

Hua-Guang ZHANG[1, 2] Xin ZHANG[3] Yan-Hong LUO[1] Jun YANG[1]

Abstract: Adaptive dynamic programming (ADP) is a novel approximate optimal control scheme, which has recently become a hot topic in the field of optimal control. As a standard approach in the field of ADP, a function approximation structure is used to approximate the solution of Hamilton-Jacobi-Bellman (HJB) equation. The approximate optimal control policy is obtained by using the offline iteration algorithm or the online update algorithm. This paper gives a review of ADP in the order of the variation on the structure of ADP scheme, the development of ADP algorithms and applications of ADP scheme, aiming to bring the reader into this novel field of optimization technology. Furthermore, the future studies are pointed out.

Keywords: Adaptive dynamic programming, neural networks (NNs), nonlinear systems, stability, optimal control

Dynamic systems are universal in nature. Stability analysis of dynamic systems has been a hot research topic for a long time, and a series of methods have been put forward. However, researchers in control theory field not only devote to guarantee the stability of the control system, but also to acquire the optimal solution. When it came to the 50 s and 60 s, because of the development of the space technology and digital computer, dynamic system optimization theory had been rapidly developed, and an important subject branch, named optimal control, emerged. It can be more and more extensively applicable in the space technology, system engineering, economic management and decision, population control, multistage process equipment optimization, and many other areas. In 1957, Bellman presented an effective tool — the dynamic programming (DP) method, which can be used for solving the optimal control problem. The Bellman principle of optimality is the key of above method, which is described as: An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. This principle can be summed up in a basic recurrence formula. When solving the multistage decision problem, we should reverse recurrence. Therefore, this principle can be applicable extensively, such as discrete systems, continuous systems, linear systems, nonlinear systems, deterministic systems, stochastic systems, and etc.

Manuscript received July 19, 2012; accepted October 29, 2012
Supported by National Basic Research Program of China (973 Program) (2009CB320601), National Natural Science Foundation of China (61104099, 61034005, 61104010), Science and Technology Research Program of The Education Department of Liaoning Province (LT2010040)
Recommended by Academician Lin HUANG
Citation: Hua-Guang Zhang, Xin Zhang, Yan-Hong Luo, Jun Yang. An overview of research on adaptive dynamic programming. *Acta Automatica Sinica*, 2013, **39**(4): 303−311
1. College of Information Science and Engineering, Northeastern University, Shenyang 110819, China 2. State Key Laboratory of Synthetical Automation for Process Industries, Shenyang 110819, China 3. College of Information and Control Engineering, China University of Petroleum, Qingdao 266580, China

Next the DP principle is introduced in two cases respectively: discrete systems and continuous systems. First discrete nonlinear systems are considered. Suppose that the system dynamic equation can be described as

$$x(k+1) = F\left(x(k), u(k), k\right), k = 0, 1, \cdots, \tag{1}$$

where $x \in \mathbf{R}^n$ represents the state vector of the system and $u \in \mathbf{R}^m$ is the control input vector. The corresponding cost function (or performance index function) has the form as

$$J\left(x(i), i\right) = \sum_{k=i}^{\infty} \gamma^{k-i} l\left(x(k), u(k), k\right), \tag{2}$$

where $x(k) = x_k$ is given, $l\left(x(k), u(k), k\right)$ is called the utility function, γ is the discount factor with $0 < \gamma \leq 1$. The objective of dynamic programming problem is to find a control sequence $u(k), k = i, i+1, \cdots$, so that the cost function in (2) is minimized.

According to Bellman principle, the minimum cost of any state from time k consists of two parts. One is the minimum cost at time k, and the other is the accumulated sum of the minimum cost from time $k+1$ to infinity, that is,

$$J^*\left(x(k)\right) = \min_{u(k)} \left\{l\left(x(k), u(k)\right) + \gamma J^*\left(x(k+1)\right)\right\}. \tag{3}$$

At the same time, the control policy $u(k)$ at time k achieves the minimum, i.e.,

$$u^*(k) = \arg\min_{u(k)} \left\{l\left(x(k), u(k)\right) + \gamma J^*\left(x(k+1)\right)\right\}. \tag{4}$$

Now we consider about the optimal control problem of nonlinear continuous-time (time-varying) dynamic (deterministic) systems, which can be described by

$$\dot{x}(t) = F\left(x(t), u(t), t\right), t \geq t_0, \tag{5}$$

where $F(x, u, t)$ is any continuous function. The objective is to choose the admissible control $u(t)$ such that the cost function (or performance index function)

$$J\left(x(t), t\right) = \int_t^{\infty} l\left(x(\tau), u(\tau)\right) \mathrm{d}\tau \tag{6}$$

achieves the minimum.

In general, we can transform the continuous-time problem into a discrete-time problem by using the discretization method, and then use the discrete dynamic programming method to find the optimal control solution. When the discretization time interval tends to be zero, both of them will tend to be consistent. With the application of the Bellman optimality principle, we can get the continuous form of DP as

$$-\frac{\partial J^*}{\partial t} = \min_{u\in U}\left\{l\left(x(t),u(t),t\right) + \left(\frac{\partial J^*}{\partial x(t)}\right)^{\mathrm{T}} F(x(t),u(t),t)\right\} = l\left(x(t),u^*(t),t\right) + \left(\frac{\partial J^*}{\partial x(t)}\right)^{\mathrm{T}} F\left(x(t),u^*(t),t\right). \quad (7)$$

From above equation, we can see that $J^*\left(x(t),t\right)$ is a first order nonlinear partial differential equation with independent variable $x(t)$, t. In mathematics, we call it as Hamilton-Jacobi-Bellman (HJB) equation.

If the system is linear and the cost function has the quadratic form with respect to the state and control input, the optimal control can be expressed as a linear feedback of the states, where the gains are obtained by solving a standard Riccati equation. However if the system is nonlinear and the cost function does not have the quadratic form with respect to the state and control input, we have to solve HJB equation to achieve the optimal control. However, it is very difficult to solve HJB equation. In addition, DP method has obvious weaknesses. With the dimension of x and u increasing, it is often computationally untenable to run true dynamic programming due to the backward numerical process required for its solution, i.e., as a result of the well known "curse of dimensionality"[1−2]. In order to overcome these weaknesses, Werbos first propose the framework of adaptive dynamic programming (ADP)[3], in which the idea is to use an approximate structure of function (such as neural network, fuzzy model, polynomial, etc.) to estimate the cost function and to solve DP problem forward-in-time.

In recent years, ADP scheme has received a widespread attention. A series of synonyms arose, for example, adaptive evaluation design[4−7], heuristic dynamic programming[8−9], neuron dynamic programming[10], adaptive dynamic programming[12] and reinforcement learning[13], etc. In 2006, National Science Foundation organized "2006 NSF Workshop and Outreach Tutorials on Approximate Dynamic Programming" seminar, where it was suggested that the kind of method is called "Adaptive/Approximate Dynamic Programming". Bertsekas et al. summarized the neuron dynamic programming[10−11]. They introduced the dynamic programming, the structure of neural network and the training algorithm, and presented many effective methods for application of neuron dynamic programming. Si et al. discussed the development of ADP scheme among inter-disciplines[14], and specially introduced the connection between DP/ADP scheme and artificial intelligence, approximation theory, control theory, operational research and statistics. In [15], Powell showed how to use ADP scheme to solve deterministic or stochastic optimization problem, and pointed out the development direction of ADP scheme. Balakrishnan et al. summarized the methods of designing feedback controller for dynamic systems by using ADP before, with the consideration of two cases i.e. for model-based systems and for model-free systems in 16. Reference [17] discussed ADP scheme from the view of whether requiring initial stability or not. Based on the research achievements of our group and the previous studies, this paper summarizes the latest development of ADP scheme.

1 Development of ADP Structure

In order to execute ADP scheme, Werbos proposed two basic structures: heuristic dynamic programming (HDP) and dual heuristic programming (DHP). The structures are shown in Fig. 1 and Fig. 2, respectively[3].

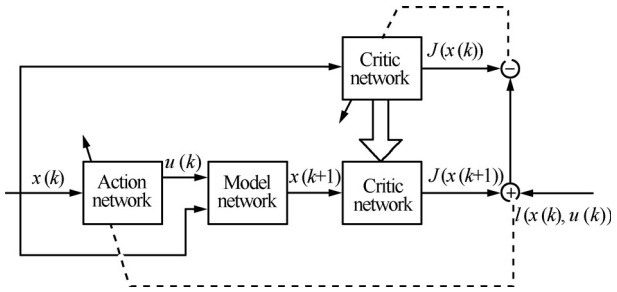

Fig. 1 The HDP structure

HDP is the most basic and widely used structure of ADP. The purpose is to estimate the system cost function. Generally three networks: critic network, action network and model network are adopted. The output of the critic network is used to estimate the cost function $J(x(k))$. The action network is used to map the relationship between state variable and the control input. The model network is used to estimate the system states for the next time. But the DHP is a method for estimating the gradient of the cost function. The definition of action network and model network is the same as the HDP, and output of the critic network is the gradient of the cost function, $\partial J(x(k))/\partial x(k)$.

Werbos further gave two other versions called "action-dependent critics", namely, ADHDP and ADDHP. The main difference from DHP and HDP is that the input of critic network is not only dependent on the system states, but including the control action. On the basis of that, Prokhorov and Wunsch presented two new structures: globalized dual heuristic programming (GDHP) and action-dependent GDHP (ADGDHP)[18], whose characteristic is that the critic network can estimate not only the cost function itself but also the gradient of the cost function. All of the above ADP structures can be used to solve the optimal control policy, but the computation burden and computa-