Neurocomputing 78 (2012) 3–13

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

# A three-network architecture for on-line learning and optimization based on adaptive dynamic programming

Haibo He [a,*], Zhen Ni [a], Jian Fu [b]

[a] Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA
[b] School of Automation, Wuhan University of Technology, Wuhan, Hubei 430070, China

ARTICLE INFO

Available online 25 August 2011

*Keywords:*
Adaptive dynamic programming
Online learning and control
Actor-critic design
Three-network architecture
Multi-state optimization
Goal representation
Reinforcement learning

ABSTRACT

In this paper, we propose a novel adaptive dynamic programming (ADP) architecture with three networks, an action network, a critic network, and a reference network, to develop internal goal-representation for online learning and optimization. Unlike the traditional ADP design normally with an action network and a critic network, our approach integrates the third network, a reference network, into the actor-critic design framework to automatically and adaptively build an internal reinforcement signal to facilitate learning and optimization overtime to accomplish goals. We present the detailed design architecture and its associated learning algorithm to explain how effective learning and optimization can be achieved in this new ADP architecture. Furthermore, we test the performance of our architecture both on the cart-pole balancing task and the triple-link inverted pendulum balancing task, which are the popular benchmarks in the community to demonstrate its learning and control performance over time.



## 1. Introduction

Learning and optimization has been a long-term focus in the machine intelligence community to understand and develop principled methodologies to replicate certain level of the brain-like general-purpose intelligence [1,2]. Over the past decades, extensive efforts have been focused on different aspects of machine intelligence research, including memory-prediction theory, embodied intelligence (EI), reinforcement learning (RL), neural dynamic programming, and many others. Recently, strong evidences from multiple disciplinary research have supported that mathematically speaking biological brain can be considered as a whole system of an intelligent controller to learn and predict to adjust actions to achieve goals [1,3]. To this end, it is widely recognized that adaptive dynamic programming (ADP) could be a core methodology to accomplish the learning and optimization capabilities for intelligent systems to approximate optimal strategy of actions over time. From the application side, ADP has also demonstrated many successful applications across a wide range of domains, such as intelligent power grid, complex system control, chess gaming, and many others [4–18].

Generally speaking, the foundation for optimization over time in stochastic processes is the Bellman equation [19], closely tied with the Cardinal utility function concept by Von Neumann. Specifically, given a system with performance cost:

$$J[\boldsymbol{x}(t),t]=\sum_{t=i}^{\infty}\gamma^{t-i}U[\boldsymbol{x}(t),u(t),t] \tag{1}$$

where $\boldsymbol{x}(t)$ is the state vector of the system, $u(t)$ is the control action, $U$ is the utility function, and $\gamma$ is a discount factor. The objective of dynamic programming is to design control sequence $u(t)$ so the cost function $J$ is minimized:

$$J^*(\boldsymbol{x}(t))=\min_{u(t)}\{U(\boldsymbol{x}(t),u(t))+\gamma J^*(\boldsymbol{x}(t+1))\} \tag{2}$$

Eq. (2) provides the foundation for implementing dynamic programming by working backward in time. For instance, universal approximators like neural networks with the backpropagation method are widely used in the community [20,21].

Existing adaptive critic design can be categorized into three major groups [5,22,23]: heuristic dynamic programming (HDP), dual heuristic dynamic programming (DHP), and globalized dual heuristic dynamic programming (GDHP). For instance, the HDP [1] was proposed with the objective of using a critic network to critique the action value in order to optimize the future cost function by using temporal differences between two consecutive estimates from the critic network [24]. This idea is essentially similar to the temporal-difference (TD) method discussed in the RL literature [25]. To overcome the limitations of scalability, DHP and GDHP were proposed [26], followed by many improvements and demonstrations of such methods [4,6,27]. The key idea of

* Corresponding author.
*E-mail addresses:* he@ele.uri.edu (H. He), ni@ele.uri.edu (Z. Ni), pigeon1387@gmail.com (J. Fu).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.05.031

DHP is to use a critic network to approximate the derivatives of the value function with respect to the states, while GDHP takes advantage of both HDP and DHP by using a critic network to approximate both the value function and its derivatives. Variations of all of these categories of ADP design have also been investigated in the community, such as the action dependent (AD) version of the aforementioned methods by taking action values as an additional input to the critic network [24].

In this paper, we aim to tackle one of the critical questions in the ADP design: how to effectively and efficiently represent the reinforcement signal to guide the intelligent system to achieve goals over time? In our approach, we propose to integrate a reference network to provide the internal reinforcement signal as a natural representation of the internal goal to facilitate machine learning. This reference network will interact with the critic network and action network in the classic ADP design to provide improved learning and optimization performance. Traditionally, in the existing ADP design, the normal way is to use a binary signal, such as either a "0" or a "−1," to represent "success" or "failure" of the system. In order to provide more informative reinforcement signals in many complex problems, different approaches have been proposed to use non-binary reinforcement signals to improve the learning performance, such as a three-value reinforcement signal (0, −0.4, and −1) for a pendulum swing up and balancing control task [24] and a quadratic reinforcement signal for the helicopter flight control [28]. Instead of hand-crafting such reinforcement signals, our objective in this paper is to seek an approach that can automatically and adaptively develop informative internal reinforcement signals to guide the intelligent system's behavior to achieve on-line learning, optimization, and control. We hope our proposed approach in this paper can provide useful suggestions to this important question regarding how to develop internal goal representations for machine intelligence research.

The rest of this paper is organized as follows. Section 2 presents the detailed three-network ADP architecture design and associated learning algorithm. The focus in this section is the reference network and critic network design, as well as their interactions to facilitate learning and optimization over time. In Section 3, detailed experimental setup and simulation analysis based on the cart-pole benchmark is presented to show the effectiveness of our approach. Following this, we present another case study in Section 4 with triple-linked inverted pendulum balancing benchmark to further demonstrate how the proposed approach performs under such a task. For both case studies, we have provided detailed comparative study of our approach with that of existing literature method. Finally, a conclusion and brief discussion regarding future research directions are discussed in Section 5.

## 2. A three-network ADP architecture

### 2.1. System architecture with internal goal representation

Fig. 1 shows the proposed ADP architecture with goal representation for learning and optimization over time. Compared to the existing ADP architectures, the key idea of our approach is to integrate another network, the *reference network*, to provide the internal reinforcement signal (internal goal representation) $s(t)$, to interact with the operation of the critic network. By introducing such a reference network to represent the system's internal goal, this architecture provides a new way to adaptively estimate the internal reinforcement signal instead of crafted by hand. This is the most important contribution of this work when compared to the existing ADP designs. From a mathematical point of view, this new architecture presents two major differences compared with that of the existing ADP designs. First, the critic network has one more additional input $s(t)$ from the reference network. Second, the optimization error function and learning in the reference network and critic network are different: The error function of the reference network is related to the primary reinforcement

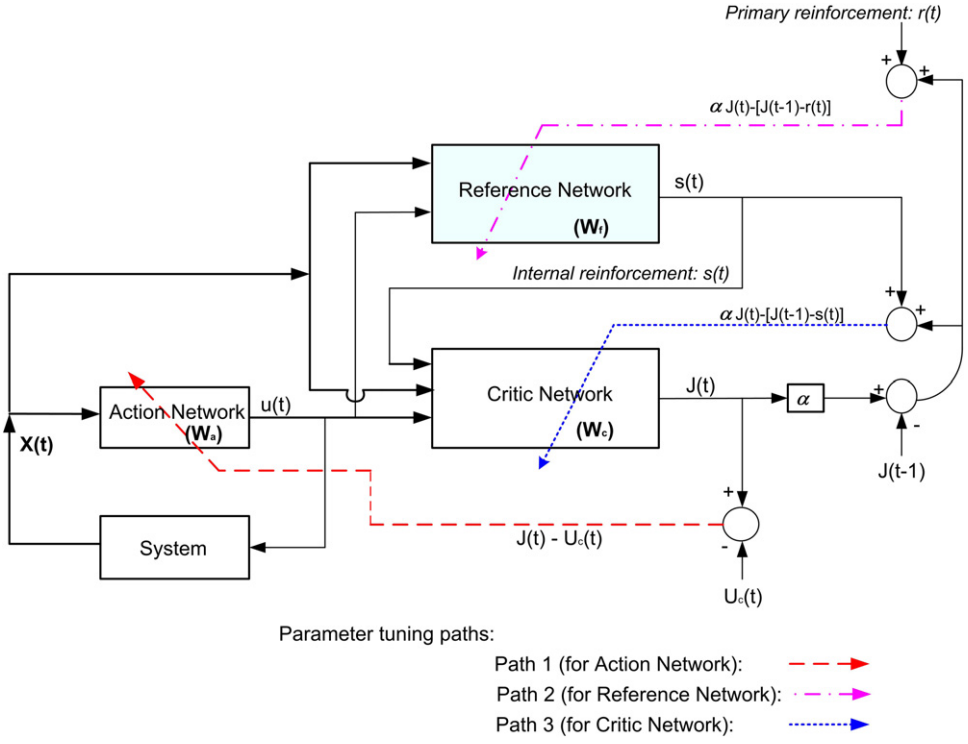


**Fig. 1.** The proposed ADP architecture with internal goal representation.