International Journal of Innovative
Computing, Information and Control
Volume **6**, Number **2**, February **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 577–590

A REINFORCEMENT LEARNING MODEL USING DETERMINISTIC STATE-ACTION SEQUENCES

Makoto Murata and Seiichi Ozawa

Graduate School of Engineering
Kobe University
1-1 Rokko-dai, Nada-ku, 657-8501 Kobe, Japan
ozawasei@kobe-u.ac.jp

Received November 2008; revised March 2009

Abstract. *This paper presents a new approach to reinforcement learning in which an optimal action policy is learned not only for primitive actions but also for deterministic state-action sequences called macro-actions. To control the exploration and exploitation of macro-actions, the temperature parameter defined by the state values and the frequency of visiting states are added to representative state-action pairs called memory items, which are stored in the long-term memory of the proposed Actor-Critic neural model. In the proposed model, no explicit form of macro-actions is defined. A macro-action is defined as a sequence of memory items with low temperature. By applying the softmax action selection to each of the memory items, an agent takes a series of actions in a deterministic way, resulting in the exploitation of a macro-action. The experimental results demonstrate that the proposed model can learn quite faster than the conventional Actor-Critic neural models in which no macro-action is introduced.*

Keywords: Reinforcement learning, Macro-action, Radial-basis function, Inductive bias

1. **Introduction.** One of the main reasons why reinforcement learning (RL) has not been a major method to build real systems is that it is generally difficult to learn an optimal policy within practical time due to its huge problem space [2, 4]. In addition, the number of training experiences is usually restricted for agents in the real world; therefore, in many RL works, an agent has to learn an optimal policy by searching a huge problem space with a relatively small number of examples. On the other hand, we can easily learn good action policies even for unknown tasks with fewer experiences. A convincing reason for this is that human can learn an action policy not only for primitive actions but also for typical sequenced of primitive actions called *macro-action*. For example, when we stand in front of a door we have never seen, first we recognize the knob, reach out the hand for it, and try to push and pull the door to open. This example implies that we often attempt to apply some typical sequences of primitive actions to solve a new task, instead of learning from all possible combinations of primitive actions. The importance of this type of studies is that learning macro-actions could provide an adaptive method to generate *inductive biases* in learning [11, 17]. Thinking about the learning process of human beings, the learning with a proper bias is a natural thought to realize intelligent autonomous systems especially when the systems are not allowed to have many experiences by interacting with real environments. Actually, several works [1, 12, 15] have demonstrated that the introduction of macro-actions allows an RL system to find a proper policy very quickly because a problem space to be searched would be effectively confined by introducing macro-actions.

There have been proposed two approaches to creating macro-actions: one is based on heuristic discovery of useful action sequences and the other is based on learning local action