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

ICIC International ©2010 ISSN 1349-4198
pp. 655–666

TWO TIME-SCALE GRADIENT APPROXIMATION ALGORITHM FOR ADAPTIVE MARKOV REWARD PROCESSES

Bing-Kun Bao, Hongsheng Xi, Baoqun Yin and Qiang Ling

Department of Automation
University of Science and Technology of China
P.O. Box 4, Hefei, 230027, P. R. China
baobk@mail.ustc.edu.cn; { xihs; bqyin; qling }@ustc.edu.cn

Received June 2008; revised December 2008

Abstract. *In this paper, we study the stochastic optimization problem of adaptive Markov reward processes parameterized by two sets of parameters, including adjustable parameters, and unknown constant parameters. As the existing algorithms do not work well for this problem, we propose a novel two time-scale gradient approximation algorithm. This new algorithm yields fast convergence, small sample path variation and low computational cost. Under some mild assumptions, we theoretically prove the convergence of the proposed algorithm, and compare it with the existing algorithms through numerical examples, which confirms its superiority.*
Keywords: Adaptive Markov reward processes, Two time-scale, Gradient approximation

1. **Introduction.** Recently, lots of interests have been caught by optimization problems [1][2][3], especially Markov reward process problem [4][5]. Markov reward process is a Markov process parameterized by a set of adjustable (i.e. tunable) parameters. The reward is associated with state transitions and the optimization goal is to maximize long-term (average or discounted) reward. Besides this set of adjustable parameters, adaptive Markov reward process has another set of parameters, which are constant but unknown, referred to as "unknown parameter" [6][7]. The key problem is to simultaneously identify these unknown parameters and seek the optimal values of adjustable parameters.

Early works on adaptive Markov reward process mostly used standard dynamic programming techniques, especially the non-stationary value iteration (NVI) scheme [7][8][9]. These techniques, however, suffer from the common weakness: when the size of the state space is large, the techniques quickly become computationally difficult, which calls "curse of dimensionality". Recently, neuro-dynamic programming techniques [10][11], such as Q-learning method [14] and TD-learning method [12][13], were introduced into this field to circumvent "curse of dimensionality". These techniques, however, cannot guarantee the performance to be improved at every iteration. Furthermore, due to the approximation error, the "optimal" policy derived from these techniques may not be the truly optimal one [15].

Simulation-based gradient stochastic approximation is promising in resolving the aforementioned issues. It is a technique relying on a parametric form of the policy (in Markov decision process) or the transition probability (in Markov reward process). It updates parameters along a single simulation sample path to search the optimal values. More specifically, this technique first estimates the performance gradient with respect to the adjustable parameters, and then adjusts the parameters along the gradient direction to