Mach Learn (2011) 85:361–391
DOI 10.1007/s10994-011-5257-4

An asymptotically optimal policy for finite support models in the multiarmed bandit problem

Junya Honda · Akimichi Takemura

Received: 4 June 2009 / Accepted: 5 June 2011 / Published online: 2 July 2011

Abstract In the multiarmed bandit problem the dilemma between exploration and exploitation in reinforcement learning is expressed as a model of a gambler playing a slot machine with multiple arms. A policy chooses an arm in each round so as to minimize the number of times that arms with suboptimal expected rewards are pulled. We propose the minimum empirical divergence (MED) policy and derive an upper bound on the finite-time regret which meets the asymptotic bound for the case of finite support models. In a setting similar to ours, Burnetas and Katehakis have already proposed an asymptotically optimal policy. However, we do not assume any knowledge of the support except for its upper and lower bounds. Furthermore, the criterion for choosing an arm, minimum empirical divergence, can be computed easily by a convex optimization technique. We confirm by simulations that the MED policy demonstrates good performance in finite time in comparison to other currently popular policies.

Keywords Bandit problems · Finite-time regret · MED policy · Convex optimization

1 Introduction

The multiarmed bandit problem is a problem based on an analogy with playing a slot machine with more than one arm or lever. Each arm has a reward distribution and the objective of a gambler is to maximize the collected sum of rewards by choosing an arm to pull for each round. There is a dilemma between exploration and exploitation: the gambler cannot

Editor: Nicolo Cesa-Bianchi.

J. Honda (✉)
Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi Chiba 277–8561, Japan
e-mail: honda@stat.t.u-tokyo.ac.jp

A. Takemura
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo-ku Tokyo 113-8656, Japan
e-mail: takemura@stat.t.u-tokyo.ac.jp