Asymptotically Optimal Policies in the Multiarmed Bandit Problem

Abstract

The multiarmed bandit problem is a model of a gambler playing a slot machine with multiple arms. It is a typical example of a dilemma between exploration and exploitation in reinforcement learning and is applied to e.g. web advertisement and network routing. In this problem a family of policies called upper confidence bound (UCB) is known to perform effectively in many models of reward distributions. However, the theoretical analysis of UCB is sometimes difficult and there are many models where it is unknown whether UCB can achieve the theoretical bound or not. To construct a policy achieving the theoretical bound, we consider a dual of UCB and develop deterministic minimum empirical divergence (DMED) policy. Whereas UCB estimates an upper bound of the expectation of each arm on some significance level, DMED estimates the upper bound of the likelihood to have larger expectation than some value for each arm. We prove that the expected loss of DMED achieves the theoretical bound for some models, such as the family of distributions on the bounded support [0,1].

---

1 Junya Honda is a Ph.D. candidate in Department of Complexity Science and Engineering at The University of Tokyo. He received his bachelor’s degree in Engineering and master’s degree in Information Science and Technology from The University of Tokyo. His research interests are in Statistical Learning Theory, Optimization, and Information and Coding Theory.