An Optimal Index Policy for the Multi-Armed Bandit Problem with Re-Initializing Bandits *

Peter Jacko
BCAM, Spain

YEQT, November 19-21, 2009

Abstract

In the multi-armed bandit problem it is assumed that the bandits are *freezing*, i.e., they maintain their state when not played. This results to be a crucial condition for proving optimality of the Gittins-index policy: "At every period play the bandit of highest Gittins index". Instead of freezing bandits, we consider *re-initializing* bandits that move back to the initial state when not played. Such dynamics is well known in reinforcement learning, where a system quits a trial and starts anew upon receiving negative feedback from the external critic, and also appears in the simplest variants of the Transmission Control Protocol implemented in the Internet. We prove that if for every re-initializing bandit there exist marginal productivity indices (which generalize Gittins indices) such that the initial state has the highest index among all the bandit's states, then the marginal-productivity-index policy is optimal. Using analogous ideas we also give a new proof of the classic problem with freezing bandits that yields insights complementary to existing proofs.

*This research has been supported by the Comunidad Autónoma de Madrid and the Universidad Carlos III de Madrid through the joint grant CCG08–UC3M/ESP-4162.