DECENTRALIZED LEARNING FOR MULTI-PLAYER MULTI-ARMED BANDITS: AN ALGORITHM FOR NEAR-LOGARITHMIC REGRET

Dileep Kalathil, Naumaan Nayyar, Rahul Jain
University of Southern California, Los Angeles, (manisser, nnayyar, rahul.jain)@usc.edu

Multi-Armed Bandits (MAB) are a classical model for learning and optimization by a single agent in an unknown environment . In this work we consider the problem of decentralized online learning with multiple players in an unknown environment which we formulate as a Multi-player Multi-armed Bandits model. In this model each player can pick among multiple arms and when a player picks an arm, it gets a random reward. We consider both i.i.d. reward model and Markovian reward model. In the i.i.d. reward model, the reward from each arm is modelled as an i.i.d. process with an unknown distribution with an unknown mean. In the Markovian reward model, each arm is modelled as a Markov chain with an unknown probability transition matrix with an unknown stationary distribution. The players have a joint objective to minimize the cost of learning called *regret* and all players should learn jointly to discover the best arms to play as a team. Since they are all trying to learn at the same time, they may collide when two or more players pick the same arm and neither of them get any reward. There is no dedicated control channel for coordination or communication among the players. Any other communication between the players is costly and will increase the regret. We propose a decentralized online learning policy called distributed Phased Exploration and Exploitation (dPEE) algorithm that achieves an expected regret that grows at most as *near-*$O(\log T)$. The motivation comes from opportunistic spectrum access by multiple secondary users in cognitive radio networks wherein they must pick among various wireless channels that look different to different users.