SHARAN VASWANI, Simon Fraser University

Global Convergence of Softmax Policy Gradient for Stochastic Bandits

Though policy gradient (PG) methods have played a vital role in the achievements of reinforcement learning (RL), the theoretical understanding of these methods is quite limited. Consequently, we focus on stochastic bandit problems (the simplest RL setting) and study the convergence of softmax policy gradient (SPG), a commonly used RL algorithm. Despite the non-concavity of the underlying objective function, recent research has leveraged the objective's smoothness and gradient domination properties to establish the convergence of SPG to an optimal policy. However, these results require setting the SPG parameters according to unknown problem-dependent quantities (e.g. the optimal action or the true reward vector in a bandit problem). We address this limitation by proposing to use SPG with exponentially decreasing step sizes. Specifically, we prove that the resulting algorithm offers similar theoretical guarantees as the state-of-the-art without requiring the knowledge of oracle-like quantities. However, using such decreasing step-sizes adversely affects the algorithm's empirical performance. Consequently, we analyze the algorithm from a different perspective and show that SPG with any constant step-size can asymptotically converge to a globally optimal policy almost surely.