CSCI 4962 Discussion 7

Siddha Kilaru

Paper 1

Auer, Cesa-Bianchi, Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," MLJ, 2002.

The K-armed bandit problem is given a set of K random variables that are independent and not necessarily independent from one another, which is the best one to sample from over time to maximize utility. An algorithm called a policy chooses the following random variable to sample from given data from prior sampling. This paper is about empirically and analytically analyzing how the regret of certain policies changes over time. Moreover, the paper uses a proof that says the regret follows a logarithmic curve over time, and the research paper was able to recreate this using different policies such as $UCB1$, $UCB2$, and ϵ-greedy. One way to extend the paper is to drop the stationarity assumption of the K-armed bandit problem, which turns it into a stochastic K-armed bandit problem. Here, there needs to be more analysis done on how regret behaves over time because the problem is much more general. This analysis can be purely analytical, or maybe some conclusions can be drawn from using an algorithm (policy) designed with the stationarity assumption in mind.

Paper 2

Mnih, Kavukcuoglu, Silver, Rusu, Veness, et al., "Human Level Control Through Deep Reinforcement Learning," Nature, 2015.

This paper is concerned with introducing deep learning methods to reinforcement learning. The paper focuses on training an agent to play classic Atari 2600 games. The paper mentions that traditional reinforcement learning methods like Q-learning have been used on these games and yielded promising results; however, combing Q-learning with deep neural networks to effectively create Deep Q-networks (DQN) performed exponentially better than the linear learning model. The design of the DQN used consisted of convolutions to a dense layer (very similar to a CNN), where the inputs were pixel data, and the outputs were controller actions. The Q-learning aspect comes into play when the weights need to be updated. Unlike traditional deep learning, there are extra steps involved in the optimization process. The results are shown in figure 3 of the paper. We can see that most of the games played by a DQN outperformed the best linear learner. Moreover, the paper mentions that the DQN was comparable

in performance to a professional human player. One way to extend this research is to study the trained DQNs and extract strategies from them. For example, we can correlate a general game state with the next most optimal play. This kind of insight can be used by human players looking to improve their scores in-game or learn patterns that were not obvious to them.

Paper 3