R7/Sanit Gupta/160100010

August 6, 2020

In this paper, the authors look at developing PAC optimal algorithms for MDPs and best arm selection in bandits. Usually, in bandits and in RL, one only looks at performance in the limit. For most popular algorithms, only their convergence in the limit has been proven. The PAC framework, on the other hand is finite time and gives bounds on the time needed to select a good enough arm or learn a good enough policy. Note that in this setting we don't care about the regret, we care only about the final arm we select. I find this to be an important paper because this gave rise to a lot of other PAC literature and is probably one of the pioneering papers in the best arm identification problem. Also, this paradigm is applicable in any real world case where we don't care about regret and have time to explore, at least in the beginning.

For the best arm selection problem in bandits, even a naive algorithm which just samples every single arm the equal number of times works and has some theoretically bounds proven on the sample complexity. The authors suggest two algorithms, one of which improves the bounds of the naive algorithm by a factor of log n.

Their first algorithm, Successive Elimination, is inspired by the setting where one knew the true means of the arms but didn't know which arm corresponds to which mean. In such a case, one could exploit the knowledge of $delta_i$'s to eliminate arms one by one in order of their difference with the optimal arm. Extending this to the setting where the true means aren't known. At any time instance, one arm is lower than the current max by more than a certain value, we can eliminate it. They keep doing this till only one arm is left. This is not PAC optimal (though this algorithm can easily be modified to become PAC), instead it just gives one the best arm with 1- δ probability.

In their second algorithm median elimination, they simply eliminate all the arms having empirical means below the empirical median after every batch of pulls. For this algorithm, the authors bound the amount by which the best true mean remaining drops at every elimination. This is enough to give PAC bounds better than the naive algorithm by log n.