R7/Ishank Juneja/16D070012

February 11, 2020

A popular view taken in Reinforcement Learning (RL) and Multi Armed Bandit (MAB) problems is one of continuing exploration and exploitation throughout the agents lifetime. Under this view, the agent learns to take better actions to maximize its expected return over a potentially infinite horizon. This seminal paper by Even-Dar and others takes a new view of the problem by introducing the probably approximately correct (PAC) framework for MABs and Markov Decision Processes (MDPs).

Under the PAC framework, the agent explores its environment for a finite horizon and at the end, produces a policy π that is ϵ-optimal with high probability. More precisely, an algorithm is called PAC-RL if for input parameters ϵ, δ and any Markov Decision Process (MDP) $\{S, A, T, R, \gamma\}$, on halting the algorithm outputs a policy π such that

$$\mathbb{P}(||V^* - V^\pi||_\infty < \epsilon) > 1 - \delta$$

A major emphasis on the paper is on applying the PAC framework to MAB problems. Unlike the full-RL problem, under the MAB setting a PAC-optimal algorithm would simply output an ϵ-optimal arm on its termination. The paper describes numerous approaches to the PAC-MAB problem starting with a "naive" algorithm that simply samples each arm a fixed number - m times and, at the end, outputs the arm with the highest empirical mean. As is intuitive, the required number of pulls m here is a function of the constraints ϵ, δ and can be computed using an application of Hoeffding's inequality. The problem with this naive method is that the total required number of pulls is $O(\frac{n}{\epsilon^2}\log(\frac{n}{\delta}))$ which can be improved upon as is seen through the subsequently presented algorithms.

Next the paper discusses an algorithm called *Successive Elimination* (SE). The algorithm starts with the assumption that the sub-optimality gaps Δ_i's are known (However their mappings to arms are unknown). In this approach, sampling is performed in slots. In each slot, arms are sampled a fixed number of times and at the end of the slot, the arm with the lowest empirical mean, up till the latest time step, is eliminated. If the largest sub-optimality gap corresponding to the worst arm were Δ_n, then each arm would be pulled at least $\frac{1}{\Delta_n^2}\log(\frac{n}{\delta})$ times, after which the empirically worst arm would be eliminated. Going forward, in the i^{th} phase, the surviving $n - i$ arms are each sampled $\left(\frac{1}{\Delta^2_{n-i}} - \frac{1}{\Delta^2_{n-i+1}}\right)\log(\frac{n}{\delta})$ assuming that the sub-optimality gaps are indexed in their ascending sequence. Adding up the number of pulls required over all the slots we find that this algorithm also has an $O(n\log(\frac{n}{\delta}))$ sample complexity. Subsequently the paper introduces an extension of SE where the knowledge of the sub-optimality gaps is not assumed. The authors show that in the modified scheme, the required sample complexity is order-wise the same as the scheme which assumes knowledge of Δ_i's. Under SE we get the same order wise complexity as the naive scheme. However, the advantage with SE is that the algorithm outputs the best arm with probability δ instead of just an ϵ-optimal arm.

The last PAC-MAB algorithm discussed in the paper is *median-elimination* (ME). Under this scheme, the algorithm throws out the worst half of the arms at each time step. Using this smarter algorithm, the authors are able to show an improved order-wise complexity of $O(\frac{n}{\epsilon^2}\log(\frac{1}{\delta}))$

A question I have is -

- No matter what PAC algorithm we use for the MAB problem, in expectation we would get linear cumulative regret (albeit with a very small slope). So is the practical usage of the PAC-MDP framework limited to the best arm identification problem?