The multi-arm bandit framework is a powerful mean for some real world applications as clinical trials, online advertisement, energy management or robotics. In the stochastic setting, an agent chooses at each time step to pull an arm among a finite number of choices. Depending on which arm that had been chosen, the agent gets a noisy reward drawn from the corresponding distribution. At a terminal stage, the agent expects to get the lowest regret possible. The usual criterion chosen to describe this regret is the expected regret: the expectation of the difference between pulling the arm with the highest mean at each time and the actual choices of the agent. The learner has to choose at each time, knowing the past, whether he has to explore other arms or exploit the rewards of the best arm discovered up to that time. The objective boils down to the exploration vs. exploitation dilemma. They are two strategies to adopt trying to solve this issue. The first is the $\epsilon$-greedy policy. In this set-up, the algorithm tries the best empirical arm with probability $1-\epsilon$ and explore other arms with probability $\epsilon$. It turns out that this strategy suffers a linear regret, which is not optimal due to the work of [Lai and Robbins]. The upper confidence bound (UCB) [Auer] utilize another strategy; optimism in the face of uncertainty. Eventually, the UCB and the KL-UCB[maillard munos] suffer a logarithmic regret – the best possible. An important issue arises then: these algorithms do well only in average. It turns out, due to an important work [Audibert] on the distribution of the regret, that the cumulative regret has a non negligible probability to be large. [Audibert2] extends this result further by proving that there is no algorithm that can get a small expected regret with exponential tails.

In some applications, the agent tries to get the best reward but have also to take into account the risk. For instance, in a clinical trial, one want to get the best treatment possible, but, in the same time, with low risks for the patient. Another field where this is common is robotics. Training robots to move may drive it to learn a movement that is optimal but that can also cause fatigue very often. There is a reward risk trade-off to solve. The first hurdle to get over is the integration of risk-aversion to the multi-arm bandit. Risk literature is very rich: from the Expected Utility theory [Newman] to the Mean-Variance setting [Markowitz]. However, there is there is no consensus over a clear definition of risk.

In this review, we will discuss the results of the integration of the risk-aversion in the multi-arm bandit framework. We will define, in the next session, the formal setup of the risk-averse multi-arm bandit. Afterwards, for each method, we will present the main algorithms devised to solve the problem before developing its theoretical guaranties.

The formal setup:

We consider the case of $K$ arms. Each arm $i \in \{1,2,\dots,K\}$ is characterized by its distribution $\nu\_i$, its mean $\mu-i$ and its variance $\sigma\_i ^ 2$. The problem is defined over a finite horizon of $T$ rounds. We denote by $X\_s^(i) \simu \nu\_i$ the s-th reward gotten from the arm $i$. All arms are

Finely

Literature

Empirical