

論文の内容の要旨

論文題目 On Thompson Sampling with Noninformative Priors in
Stochastic Multi-Armed Bandits

(確率的多腕バンディットにおける無情報事前分布を用いたトンプソン抽出について)

氏 名 李 鍾瑛

We human beings always make a decision on a daily basis. Such a decision would be a trivial one, such as what to eat for lunch, or an important business one that determines the future of the company. This is why decision-making modeling has been proposed and analyzed in several research contexts, such as business, political psychology, statistics, and computer science. In particular decision-making scenarios, one can observe all possible outcomes regardless of their decisions. However, making decisions is not always straightforward, as we often have to take action based on imperfect observations. This is because we usually only observe the consequences of our actions, which is known as partial feedback.

In the sequential scenario, an agent aims to achieve their goal by exploring unknown actions to decrease their uncertainty or exploiting well-known actions that are expected to be appropriate to achieve their goal. The Multi-Armed Bandit (MAB) problem is a classical model that exemplifies such a trade-off between exploration and exploitation in sequential decision-making.

This thesis studies the stochastic setting, where an agent observes an outcome (or reward) every time they select an action (or arm) generated from the underlying arm distribution. To solve such problems efficiently, the agent would pose an assumption on the reward distribution with their prior knowledge. For example, one could assume that rewards are generated from the Bernoulli distribution if rewards are binary.

In stochastic MAB problems, two objectives have been mainly considered: (1) regret minimization where the agent aims to maximize their rewards during the time horizon, which is a problem of balancing the trade-off between exploration and exploitation. (2) best-arm identification where the agent aims to identify the best (optimal) arm with limited resources, which is a problem of pure exploration.

Thompson sampling is a randomized probability matching policy that solves the stochastic MAB problem by playing arms according to the posterior probability of being optimal. From its randomization nature, Thompson sampling makes the balance naturally between exploration and exploitation and has shown excellent performance in practice. However, despite its effectiveness, Thompson sampling was not considered an attractive choice because of its lack of theoretical understanding for a long time. In the last decade, the theoretical analyses of Thompson sampling have been conducted, and now it is considered one of the main approaches to the stochastic MAB problem.

In much literature on Thompson sampling in the regret minimization problem, the underlying distribution was assumed to belong to the single parameter exponential family, such as the Bernoulli distribution, or to be a light-tailed distribution, where its tail probability decreases exponentially. This is because the sub-Gaussian noise is widely observable in many problems, and it is easy to control the probability of extreme events. In practice, however, multi-parameter or heavy-tailed distributions have also been widely adopted to analyze stochastic systems in several research fields, such as economics and signal processing. Nevertheless, Thompson sampling has rarely been investigated under such distributions, even in the parametric stochastic MAB problem, which is a classical problem.

The first part of this thesis aims to deepen the theoretical understanding of

Thompson sampling in such distributions from a problem-dependent view, where we show the problem-dependent optimality of Thompson sampling depends on the choice of (non-informative) priors. More precisely, we first provide a theoretical analysis of Thompson sampling in the uniform distribution with unknown support, which is a two-dimensional parametric distribution and does not belong to the exponential family. We further propose a variant of Thompson sampling that could maintain optimality under one-to-one reparameterization, as our analysis shows the optimality of Thompson sampling further depends on the parameterization of distributions. We then extend our analysis to the Pareto distribution, which is a heavy-tailed distribution parameterized by two unknown parameters.

It is worth noting that online ad allocation is one main application of bandit algorithms, and the Pareto distribution is commonly observed in the analysis of the internet and web.

The second part of this thesis focuses on the best-arm identification problem with Thompson sampling. Despite its efficient exploration in the regret minimization problem, direct use of Thompson sampling in pure exploration cannot make the optimal algorithm. This is because Thompson sampling plays a suboptimal arm a log-order times, which induces a bias in the number of playing arms. To mitigate this issue, we concentrate our efforts on leveraging the inherent randomization characteristic of Thompson Sampling, integrating it with deterministic algorithms to construct a more balanced and effective exploration strategy.

In summary, this thesis aims to extend the theoretical understanding of Thompson sampling in stochastic MAB problems, placing a specific emphasis on guiding the selection of priors in general models. Such extensions are not only for mathematical interests, but they will also be helpful to practitioners who want to solve sequential decision-making problems through the application of easy-to-implement algorithms that guarantee excellent performance in both theory and practice. The insights and results elucidated in this thesis significantly augment our understanding of Thompson Sampling, offering enhanced and efficient solutions to MAB problems.