

論文の内容の要旨

論文題目

A Study on Stochastic Combinatorial Bandit Problems
with Limited Observation
(限られた観測に基づく確率的組合せバンディットの研究)

氏 名 黒木 祐子

Decision making under uncertainty is common in prudent weighing of outcomes or utility because we human beings are constantly exposed to risk or ambiguity situations. In the face of uncertainty in our lives, we aim to take the best possible decision based on the past experience and conjecture. Modeling of decision making under uncertainty and its analysis have been investigated in diverse research fields such as computer science, statistics, psychology, and neuroscience.

The problem of *stochastic multi-armed bandits* is one of the classical decision making models, and it lies at the intersection of statistics and machine learning. The stochastic multi-armed bandit problem characterizes the trade-off between exploration and exploitation in stochastic environments; in this problem, an agent chooses one action (a. k. a. arm) from K given arms at each step, and obtains a reward sampled from an unknown probability distribution associated with the chosen arm.

In many real-world scenarios, our decisions are often characterized by a combinatorial and complex structure. For example, possible actions in real-world systems may be a subset of keywords in online advertisements, assignments of tasks

to workers in crowdsourcing, or paths in communication networks. The problem of stochastic combinatorial bandits is a generalization of the multi-armed bandits, which can deal with such combinatorial actions; a subset of arms (a.k.a. super arm) is an action in this model, while each single arm is an action in the multi-armed bandit problem.

In this thesis, we study *stochastic combinatorial bandit* problems from an algorithmic perspective. Motivated by real-world applications, we develop theory and algorithms for challenging settings, in which only limited feedback is given to an agent. In the combinatorial bandit literature, a large amount of work has studied the semi-bandit setting, in which an agent can directly observe a random feedback from an individual arm at each sampling trial. However, such a sampling procedure for individual arms is costly or often impossible due to the privacy issues or system constraints. To overcome this critical limitation of existing studies, we aim to propose statistically and computationally efficient methods based on only limited observation. To this end, we address several approximation algorithms for NP-hard problems and bandit algorithms that can deal with an aggregated feedback from a super arm.

This thesis focuses on the combinatorial bandit problems in the *pure exploration* setting; in this setting, an agent aims to identify the optimal super arm with the highest expected reward. The learning performance is evaluated by the number of samples required by her learning algorithm, i.e., the sample complexity, rather than the cumulative rewards. The main body of this thesis consists of four parts, which corresponds to Chapters 3–6. The first part is devoted to a study of pure exploration for top- k arms with *full-bandit feedback*, where underlying combinatorial structures are simply size- k subsets but available feedback is limited to a noisy evaluation for a sampled subset of arms, i.e., full-bandit feedback. The second part is devoted to a study of general combinatorial pure exploration with full-bandit feedback, in which possible combinatorial structures are size- k subsets, matchings, and paths. The third part is devoted to a study of combinatorial pure exploration over graphs with full-bandit feedback, which aims to identify a dense component in a network only together with a noisy evaluation for a sampled subgraph. The fourth part is devoted to a study of general combinatorial pure exploration with *partial-linear feedback*, which aims to develop a novel algorithmic framework for general (possibly nonlinear) reward

functions and combinatorial structures including size- k subsets, matchings, and paths. Partial-linear feedback ranges from semi-bandit and full-bandit feedback; in this feedback model, the agent only observes a linear combination of rewards of the chosen super arm at each pull.

In Chapter 3, we first investigate a novel problem of pure exploration for top- k arms with full-bandit feedback. Although our problem can be regarded as an instance of linear bandit problems, a naive approach using linear bandit algorithms is computationally infeasible to the problem instance in the combinatorial setting, since the number of possible actions K is exponential with respect to the number of arms in the subset. To cope with this problem, we design a polynomial-time approximation algorithm for the 0-1 quadratic programming problem arising in confidence ellipsoid maximization. Based on our approximation algorithm, we propose a bandit algorithm whose time complexity is $O(\log K)$, thereby achieving an exponential speedup over linear bandit algorithms. We also provide an upper bound on the sample complexity that is worst-case optimal. Our experiments on large-scale crowdsourcing datasets with more than exponentially many possible actions demonstrate the superiority of our algorithm in terms of both the computation time and the sample complexity.

In Chapter 4, we study combinatorial pure exploration with full-bandit feedback for general combinatorial structures such as matroids, matchings, and s - t paths. We devise a polynomial-time adaptive algorithm whose sample complexity has mild dependence on the *minimal gap* between the optimal value and the second optimal value. We show that the sample complexity matches (within a logarithmic factor) the information theoretic lower bound for a family of instances. We conduct a series of experiments on the matching and top- k instances and demonstrate that (a) the proposed algorithm scales to combinatorial instances while existing linear bandit algorithms require a prohibitive amount of time; (b) the number of samples required by the proposed algorithm is not sensitive to the value of the minimum gap.

In Chapter 5, we focus on a specific instance of combinatorial pure exploration with linear feedback over graphs with applications to *dense subgraph discovery under uncertainty*. We introduce a novel learning problem for online dense subgraph discovery, in which an agent queries subsets of edges rather than single edges

and observes a noisy sum of edge weights in a queried subset. For this problem, we propose a polynomial-time algorithm that obtains a nearly-optimal solution with high probability. For the proposed algorithm, we provide an upper bound of the number of samples that the algorithm requires. Moreover, to deal with large-sized graphs, we design a more scalable algorithm that maximizes the quality of solution within a fixed number of queries. Computational experiments using real-world graphs demonstrate the effectiveness of our algorithms.

In Chapter 6, we study a novel model of general combinatorial pure exploration with partial-linear feedback, which includes all the problems addressed in Chapters 3–5 as special cases and finds various applications such as online ranking in recommendation systems and crowdsourcing. The model of partial-linear feedback fits in such real-world applications since it may happen that we cannot always observe outcomes from some of the chosen arms due to privacy concern or system constraints. We propose a polynomial-time algorithmic framework for this general problem, and provide a sample complexity analysis.

To summarize, we investigate the learnability of the combinatorial pure exploration problems with limited feedback and develop polynomial-time algorithms. Our results provide novel insights into online decision making problems with combinatorial action spaces and combinatorial optimization under uncertainty for incomplete inputs.