A Behavioral Model for Exploration vs. Exploitation: Theoretical Framework and Experimental Evidence

Abstract: The exploration-exploitation trade-off is a fundamental concept arising when the decision maker needs to make repeated choices, whose rewards are unknown *in priori*. This paper delves into how humans navigate this trade-off through the lens of the multi-armed bandit (MAB) problem. While much literature primarily focuses on the design and performance of (near) optimal algorithms for different variants of MAB, few studies are devoted to understanding the DM's exploration-exploitation trade-offs behaviorally. In this regard, our research question is to investigate a behavioral model to characterize the exploration-exploitation trade-off of human beings.

Inspired by both behavioral economics and online learning literature, we introduce a novel family of behavioral policies called Myopic Quantal Choice (MQC). In an MQC, the probabilities of choosing the arms are based on the quantal choice model. It is a dynamic adaptation of quantal choice models, deriving anticipated utilities directly from past rewards. MQC offers a simple method to describe the arm selection process, and yet is rich enough to quantify the exploration-exploitation trade-off through a "shrinkage rate of exploration." We posit that there is considerable heterogeneity among individuals in how they dynamically adjust their degree of exploration, and encapsulate these behaviors with a parameter to represent the shrinkage rate of exploration.

Through both non-asymptotic and asymptotic analysis, we show that MQC admits intuitive properties that match the qualitative patterns of the laboratory experiment data. Particularly, MQC always converges to the optimal arm, thus capturing the "learning" effect. The main results of this paper can be summarized from two perspectives: the theoretical analysis of MQC, and the analysis of the human behavioral data.

1. We characterize MQC's regret in the asymptotic regime and demonstrate the effects of "over-" and "under" exploration: Over-exploration with a too small shrinkage rate

parameter results in gradual deterioration of the lower bound of regret, while under-exploration with a too small rate parameter leads to sudden deterioration.

2. We conduct behavioral lab experiments to investigate human beings' behavior in MAB and provide an analysis of the behavioral data. Analysis of laboratory experiment data reveals that the MQC model excels in predictive power compared to other behavioral models. Insights from the asymptotic regime also extend to the finite horizon experiments. Particularly, when we fit the MQC model to the data, a prevalent tendency toward over-exploration among subjects becomes evident.

Based on theoretical analysis and numerical experiments, we believe the significance of MQC is three-fold.

1. First, it provides a conceptual framework to understand the DM's exploration-exploitation trade-off and capture the dynamic adjustment.

2. Second, it can be used as a structural estimation tool to learn from realized actions and rewards how much human beings are "under" or "over" exploring.

3. Finally, it can be extended to a more general form and provide a more principled justification of the exploration factor.