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

**Jingying Ding**
Shanghai University of Finance and Economics
318 Wuchuan Rd, Shanghai, China 200437
dingjingying@mail.shufe.edu.cn

**Yifan Feng**
National University of Singapore
Singapore 119245
yifan.feng@nus.edu.sg

**Ying Rong**
Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai, China 200240
yrong@sjtu.edu.cn

## Abstract

How do people navigate the exploration-exploitation (EE) trade-off when making repeated choices with unknown rewards? We study this question through the lens of multi-armed bandit problems and introduce a novel behavioral model, *Quantal Choice with Adaptive Reduction of Exploration* (QCARE). It generalizes Thompson Sampling, allowing for a principled way to quantify the EE trade-off and reflect human decision-making patterns. The model adaptively reduces exploration as information accumulates, with the reduction rate serving as a parameter to quantify the EE trade-off dynamics. We theoretically analyze how varying reduction rates influence decision quality, shedding light on the effects of "over-exploration" and "under-exploration." Empirically, we validate QCARE through experiments collecting behavioral data from human participants. QCARE not only captures critical behavioral patterns in the EE trade-off but also outperforms alternative models in predictive power. Our analysis reveals a behavioral tendency toward over-exploration.

In many business settings, decision-makers repeatedly choose among options with unknown rewards, learning from experiencee.g., consumers picking unfamiliar brands or managers selecting suppliers without known reliability. The decision-making process can be very intricate. In particular, it necessitates a delicate balance between *exploration* – trying different options to learn about potential rewards – and *exploitation* – exploiting the best-known options to collect rewards. The exploration-exploitation (EE) trade-off is usually modeled by the multi-armed bandit problem (MAB). In this problem, the decision maker faces several slot machines (or "arms"), each with an unknown reward probability distribution. The decision maker's goal is to maximize their reward over time by choosing which arms to pull. This problem has drawn attention from many communities such as economics, computer science, and management science. Significant advancements in algorithmic solutions have been made, such as the Gittens Index policy, Thompson Sampling, and Upper Confidence Bound methods. (See more details in the literature review.)

However, the algorithms designed to maximize rewards for MAB problems may behave very differently from *human behavior* when they face similar problems. For example, [2] demonstrated that Gittins index-based policies may not describe human behavioral data as well as seemingly naive

*Full version: https://arxiv.org/abs/2207.01028.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

models, such as hot-hand and exponential smoothing policies. Given that many repeated decisions in business practices are not dictated by algorithms but rather managed by human beings, our paper is driven by the following questions: How do human decision-makers balance the trade-off between exploration and exploitation? How does varying the balance between exploration and exploitation impact decision quality?

**Summary of Results and Contributions**

**Modeling contributions.** In this paper, we introduce a novel model that bridges algorithmic and behavioral modeling to shed light on the dynamic interplay between exploration and exploitation. Our model leverages concepts from the quantal choice framework, where the decision-maker relies on a (randomized) score system and chooses the arm with the highest attraction score. On a high level, the score is arm- and history- dependent and takes the following form:

$$\text{SCORE} = \text{HISTORICAL PERFORMANCE} + \text{WEIGHT} \times \text{RANDOM SHOCK}.$$

We relay to the full paper for a precise formula. This score system explicitly encodes the trade-off between exploration and exploitation: The first is purely determined by the historical performance of the arm. It signifies exploitation as it leads the decision maker to choose the arm with the best historical performance. The second component consists of random shocks. It enables exploration by trying arms that have not performed well historically. The weight term thus controls the balance between exploration and exploitation.

An important feature of our model is that the weight will shrink with the number of times the arm has been pulled. In other words, exploitation will be emphasized over exploration over time as the decision maker's experience accumulates. This feature is inspired by qualitative evidence from experiments. It is also why we term of our model as *Quantal Choice with Adaptive Reduction of Exploration* (QCARE).

We parameterize the decision maker's EE trade-off dynamics into a structural parameter. We refer to it as the *reduction rate of exploration* and denote it as $\alpha$. The larger the value of $\alpha$, the faster the reduction rate of exploration, and therefore exploitation dominates exploration more quickly. The specific form of the weight draws inspiration from the online learning literature: QCARE can be viewed as a generalization of the well-celebrated Thompson Sampling (TS) method. When $\alpha = 0.5$, our model reduces to Gaussian TS. When $\alpha < 0.5$ (resp. $\alpha > 0.5$), our model captures a policy that explores more (resp. less) aggressively than TS. Empirically, the value of $\alpha$ can then be estimated from behavioral data.

The marriage between online learning and behavioral modeling offers our model a few advantages, which we summarize below.

- First, it captures the learning effect for dynamic choices in a *parsimonious* way. When it comes to dynamic choice models, the traditional approach determines choice probabilities in a "subgame-perfect" manner (e.g., 3, 4, 1). However, the computation of value functions suffers from the curse of dimensionality. In comparison, the scores in our model admit *myopic* forms and thus allow simplicity for both theoretical analysis and empirical estimation.
- Second, it provides an interpretable yet *principled* quantification of the EE trade-off. We theoretically characterize how EE trade-off affects decision quality, thus justifying key concepts such as over- and under- exploration. (See our summary of technical contributions below.)
- Third, the QCARE policy family includes not only (asymptotically) optimal policies such as Thompson Sampling, but also suboptimal ones due to over- and under- exploration. This allows QCARE to capture choice behavior with potentially bounded rationality. The benefit of this flexibility is reflected by how it displays better empirical performance on behavioral data, as well as unlocking novel behavioral patterns. (See our summary of experimental and empirical contributions below.)

**Theoretical contributions.** We study both non-asymptotic and asymptotic properties of QCARE. We show that QCARE displays comparative statistics properties that are qualitatively consistent with our lab evidence. In addition, we show that all $\alpha > 0$ enables the decision maker to converge to the optimal arm in the long run. It suggests that every $\alpha > 0$ is "plausible" – at least when $T$ is large – since they all correspond to long-run-average optimal policies.

Besides intuitive qualitative properties, QCARE offers a principled way to quantify the EE trade-off using $\alpha$. We develop an asymptotic theory regarding how different alpha values lead to different decision qualities, measured by regret.

- When $\alpha = 0.5$, QCARE achieves the optimal regret order of $O(\sqrt{T})$. In other words, in the asymptotic regime where $T = \infty$ represents the "optimal" balance of exploration vs. exploitation.
- When $\alpha < 0.5$, the regret order gradually deteriorates to $\Omega(T^{1-\alpha})$.
- When $\alpha > 0.5$, the regret order worsens drastically to $\Omega(T^{1-\varepsilon})$ for every $\varepsilon > 0$.

The analysis above characterizes the effects of "over" and "under" exploration asymptotically, which we illustrate in Figure 1.

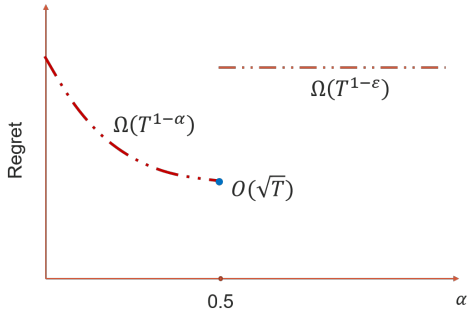


Figure 1: Asymptotic Regret of QCARE as a function of $\alpha$

We extend our analysis in two dimensions. First, to deepen the theoretical understanding of the asymptotic results above, we study a general family of Markovian MAB policies and identify conditions under which the same asymptotic pattern of over- and under- exploration emerge. Second, we conduct extensive numerical studies and find that the insight from asymptotic analysis generalizes to the non-asymptotic setting where $T$ is fixed to be a moderate value.

**Experimental and Empirical contributions.** We conduct experiments to collect behavioral data regarding how real human beings make decisions in the MAB problem. QCARE displays strong capabilities of capturing human behavior, especially in terms of out-of-sample prediction power compared to other models drawn from different streams of literature. Through our analysis, we discover an interesting behavioral pattern: people tend to "over-explore" in the sense that they settle at leading arm more slowly than the expected-reward-maximizing rate. While such bias could be partly rationalized by the risk aversion of the participants, their behavior of not choosing the leading arm appears to reflect a mix of random behavioral errors (in the same spirit of 5) layered on top of the "intrinsic" exploration required for dynamic learning.

### Analysis of A More General Framework

We study the decision dynamics and regret of QCARE by embedding it in and characterizing a general class of Markovian MAB policies. We then verify which of these conditions QCARE satisfies under different values of $\alpha$. Unlike traditional analyses that focus only on order-optimal policies, our framework applies to a wide range of behavioral policies, including those that are over- or under-exploring. As such, we believe our analysis makes a theoretical contribution to the analysis of MAB policies in its own right.

**Setup and basic properties.** Let us first describe the family of MAB policies our analysis extends to. Let $k_i(t)$ and $\hat{\mu}_i(t)$ represent the pull count and empirical reward of arm $i$ up to period $t-1$. We consider MAB policies that only depend on $\boldsymbol{S}(t) := (k_1(t), \ldots, k_N(t), \hat{\mu}_1(t), \ldots, \hat{\mu}_N(t))$ as a sufficient statistic, which takes values in $\mathcal{S} := \mathbb{Z}_+^N \times [0,1]^N$. We also index a typical state as $S = (\kappa, u)$, where $\kappa := (\kappa_1, \ldots, \kappa_N) \in \mathbb{Z}_+^N$ and $u := (u_1, \ldots, u_N) \in [0,1]^N$ are scalar vectors. In this way, an admissible policy can be represented by a probability function $Q = (Q_1, \ldots, Q_N) : \mathcal{S} \to \Delta_N$, where $Q_i(S) = \Pr(a(t) = i \mid \boldsymbol{S}(t) = S)$ is the probability to pull arm $i$ given state $S$.

With a slight abuse of notation, we also find it convenient to write $Q_i(S)$ as $Q_i(S) = Q_i(S^i; S^{-i})$, where $S^i = (\kappa_i, u_i)$ and $S^{-i} = (\kappa_1, \ldots, \kappa_{i-1}, \kappa_{i+1}, \kappa_N, u_1, \ldots, u_{i-1}, u_{i+1}, ..., u_N)$ are the state values for arm $i$ and the rest of the arms, respectively. The family of MAB policies we consider is equivalent to the *Sequentially Randomized Markov Experiments* studied by [6]. QCARE belongs to this family, and we refer the reader to [6] for many other members of this family.[2]

**Conditions for "over" and "under" exploration.** Let us now proceed to two key novel measures we use to quantify "over" and under" exploration. We start with the one corresponding to over-exploration.

**Definition.** *Let $\alpha \in (0, 0.5)$ be fixed. The probability function $Q$ is $\alpha$-exploratory if there exists constants $\varepsilon > 0$, $\delta \in (0, 1)$ and a sequence of numbers $\{\Delta_1, \Delta_2, \ldots\}$,where $\Delta_T = \Omega(T^{-\alpha})$, such that for all sufficiently large $T$, $u_1 \leq \Delta_T$, and $\kappa$ such that $\sum_{i>1} \kappa_i < \delta T$, it holds that*

$$1 - Q_1 (\kappa_1, \kappa_2, ..., \kappa_N, u_1, 0, 0, ..., 0) \geq \varepsilon.$$

Roughly speaking, a policy with an $\alpha$-exploratory function $Q$ allows situations where even all arms have been pulled $\Omega(T)$ times and their empirical reward differences on the order of $\Omega(T^{-\alpha})$, the arm-pulling probabilities do not concentrate on the leading arm. In other words, an $\alpha$-exploratory function $Q$ with a small $\alpha$ is a sign of heavy exploration. Intuitively, if the exploration is too heavy, then the policy may fail to concentrate on any particular arm despite overwhelming evidence that a certain arm the the best. In the result below, we formally characterize the consequence of such "over-exploration."

**Theorem.** *Let $\alpha \in (0, 0.5)$ and $N > 1$ be fixed. If the probability function $Q$ is $\alpha$-exploratory, then the regret satisfies $\mathcal{R}(T) = \Omega(T^{1-\alpha})$.*

The next property provides a quantitative measure of how a policy could potentially "under explore."

**Definition.** *Let $\alpha \in (0.5, +\infty)$ and $N > 1$ be fixed. The function $Q$ is $\alpha$-irreversible if for any following list of quantities: (i) arbitrarily small constant $\delta > 0$, (ii) arm $j > 1$, (iii) $u \in [0, 1]^N$ satisfying $u_j - u_1 = \Omega(1)$, and (iv) $\kappa \in \mathbb{Z}_+^N$ satisfying $\kappa_1, \kappa_j \geq (\log T)^{\frac{1}{2\alpha - \delta}}$ for sufficiently large $T$, it holds that $Q_1(\kappa, u) \leq \frac{(\log T)^{\frac{1}{2\alpha-\delta}}}{T}$.*

Roughly speaking, a policy with an $\alpha$-irreversible function $Q$ allows situations where there are two arms where both are pulled only $(\log T)^{\frac{1}{2\alpha-\delta}}$ times, while the worse-performing arm's pulling probability already diminishes on the order of $O\left(\frac{(\log T)^{\frac{1}{2\alpha-\delta}}}{T}\right)$. This loosely translates to the probability of pulling the non-leading arms decaying super-exponentially fast with the pull count. Hence that corresponds to a sign of aggressive concentration (or lack of exploration). Intuitively, if the concentration is too aggressive, then the policy may be mistakenly stuck in the inferior arm in a "bad" event where it performs better empirically during the initial periods. This intuition is also where the name "irreversibility" comes from. In the result below, we formally characterize the performance consequence for "under-exploration."

**Theorem.** *Suppose the probability function $Q$ is unradical, i.e., for $\kappa \in \mathbb{Z}_+^N$ and $u \in [0, 1]^N$ such that $u_1 \leq \min\{u_2, \ldots, u_N\}$, $Q_1(\kappa, u) \leq \frac{1}{2}$. Then even when $N = 2$, for every policy with an $\alpha$-irreversible function $Q$, the regret satisfies $\mathcal{R}(T) = \Omega(T^{1-o(1)})$.*

The theorem reveals that without sufficient exploration, the aforementioned "bad" event happens with probability $\Omega(T^{-o(1)})$ and the regret conditional that bad event is on the order of $\Omega(T)$. That makes the overall regret deteriorate almost linear in $T$.

[2]We assume that the probability function cannot "cheat" by depending on the arm identity information. Formally, this is by assuming that it is invariant to relabeling. That is, for every permutation $\sigma : [N] \to [N]$, $i \in [N]$, and state $S = (\kappa_1, \ldots, \kappa_N, u_1, \ldots, u_N) \in \mathcal{S}$, $Q_i(\kappa_{\sigma(1)}, \ldots, \kappa_{\sigma(N)}, u_{\sigma(1)}, \ldots, u_{\sigma(N)}) = Q_{\sigma(i)}(\kappa_1, \ldots, \kappa_N, u_1, \ldots, u_N)$.

## References


[1] Tülin Erdem and Michael P Keane. Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science*, 15(1):1–20, 1996.

[2] Noah Gans, George Knox, and Rachel Croson. Simple models of discrete choice and their performance in bandit experiments. *Manufacturing & Service Operations Management*, 9(4):383–408, 2007.

[3] John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

[4] John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033, 1987.

[5] Xuanming Su. Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management*, 10(4):566–589, 2008.

[6] Kuang Xu and Stefan Wager. Weak signal asymptotics for sequentially randomized experiments. *Management Science*, 70(10):7024–7041, 2024.