Learning and Earning Under Noise and Uncertainty

Su Jia
Tepper School of Business
Carnegie Mellon University

Dissertation Committee:
R. Ravi (Chair)
Andrew Li
Alan Scheller-Wolf
Sridhar Tayur

First version: Apr 29, 2022.
Updated: Nov 1, 2024

Introduction

Sequential decision-making under uncertainty is central to a range of operations and marketing problems. In the face of an unknown environment, the decision-maker needs to strike a balance between learning the known environment ("learning") and selecting nearly optimal decisions ("earning"). For example, consider pricing a new product. If the retailer knew the demand function (i.e., the relationship between mean demand as a function of the price), they would simply select the revenue-maximizing price. However, the demand function may not be available in practice, so the seller needs to experiment with different prices to *learn* the demand function, and they make pricing decisions to *earn* revenues based on the observed data.

The learning-earning trade-off can be captured by the *Multi-Armed Bandits* (MAB) framework, which has recently attracted significant attention from a range of communities, including machine learning, operations research and marketing. While most of the fundamental problems have been theoretically well-understood, these algorithms have been rarely deployed in practice. This thesis serves as a preliminary step towards filling this gap between practice and theory.

We address three challenging aspects for balancing learning and earning:
(i) **Monotonicity constraint:** Markdown Pricing Under Unknown Demand (Section 1),
(ii) **Combinatorial structures**: Optimal Decision Tree with Noise (Section 2), and
(iii) **Large action space:** Short-lived High-volume Bandits (Section 3).
These challenges arise from fundamental areas in marketing, including pricing, survey design, and content recommendation. We offer theoretical insights with performance guarantees and empirical evidence demonstrating why our methodology outperform existing approaches.

1 Monotonicity Constraint

Dynamic pricing under unknown demand has been theoretically well-understood, usually under the framework of multi-armed bandits. But in practice, these bandit-based policies are rarely deployed by real-world retailers, largely because oscillating prices may cause customer dissatisfaction. For example, Luca and Reshef [2021] found that a "1% price increase (in menu prices) leads to a 3% to 5% decrease in online ratings on average".

This motivates us to consider a dynamic **markdown** pricing problem, where the price sequence is **required** to be non-increasing. While both (a) markdown pricing under *known* demand and (a) *non*-markdown pricing under unknown demand have been well-understood, little is known for markdown pricing under a unknown demand function. In particular, a basic question remains open:

What is the minimax optimal regret for markdown pricing under unknown demand?

And even further, for various families of demand functions,

Is the optimal regret for markdown pricing higher than that of non-markdown pricing?

Thus motivated, we formulate a continuum-armed bandit problem where the arm sequence is required to be non-increasing. Each arm corresponds to a price, and the reward function is the product of the price and the demand function. The unknown demand function comes from a given class, which may be non-parametric [Jia et al., 2021] or parametric Jia et al. [2023a].

	Due to	Family	Bdd. r''	Monotone	Regret
Nonparam.	Kleinberg [2005]	Lipschitz	no	no	$\tilde{\Theta}(n^{2/3})$
	Babaioff et al. [2015]	MHR	no	yes	$\tilde{O}(n^{3/4})$
	Jia et al. [2021], Chen [2021]	Lipschitz	no	yes	$\Theta(n)$
	Jia et al. [2021],Chen [2021]	unimodal	no	yes	$\tilde{\Theta}(n^{3/4})$
	Jia et al. [2021]	unimodal	yes	yes	$\tilde{\Theta}(n^{5/7})$
Param.	Broder and Rusmevichientong [2012]	k-crossing ($k \geq 1$)	yes	no	$\tilde{\Theta}(\sqrt{n})$
	Jia et al. [2023a]	k-crossing ($k \geq 1$)	yes	yes	$\tilde{\Theta}(n^{k/k+1})$
	Broder and Rusmevichientong [2012]	0-crossing	yes	no	$\Theta(\log n)$
	Jia et al. [2023a]	0-crossing	yes	yes	$\Theta(\log^2 n)$

Table 1: **Minimax Regret, With and Without Monotonicity.** Our results (colored **blue**) are **optimal** up to logarithmic (or constant) terms. Here, r'' is the second derivative of the revenue function r. We introduced the term "k-crossing" which roughly means that any two curves in the family intersect k times at most. In particular, 0-crossing means any two curves are disjoint. Non-parametric results still hold if the initial inventory m is finite, with n replaced by $\max\{m, n\}$.

We provide a **complete** settlement of the problem by presenting novel learning policies with minimax optimal regret, summarized in Table 1. Notably, most of these tight bounds are asymptotically higher than their unconstrained counterparts (expect the 1-crossing family, e.g., linear demand function), underscoring the impact of monotonicity.

2 Optimal Decision Tree Under Noisy Outcomes

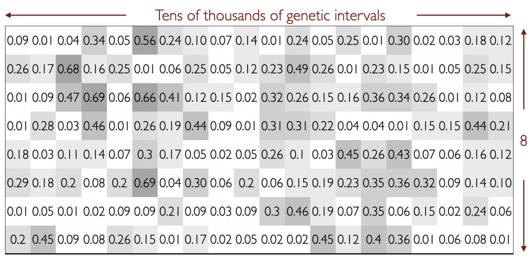
From Spotify to Netflix, we experience personalization in nearly every aspect of daily life, and consumers now expect this same level of tailored engagement from companies of all sizes. One effective strategy for personalizing service for new users involves categorizing them into typical user types and identifying their type through responses to survey questions.

Combinatorial structures in the these problems (such as those enabling binary search) can usually be leveraged to speed up learning. This is encapsulated by the classical *Optimal Decision Tree* (ODT) problem. Given a set of n *tests*, a set of m *hypotheses*, and an $m \times n$ table encoding the outcome for each pair of test and hypothesis, we aim to find a low-cost testing procedure (*decision tree*) that identifies the true hypothesis. The deterministic-outcome setting has been extensively studied. However, in many applications, the outcomes may be uncertain, which renders the ideas in the deterministic setting invalid.

Adaptivity	**Assumption**	**Apxn. ratio**	**APX-hardness**
non-adaptive	none	$\log m$	$\log m$
adaptive	**Few Unknowns:** $\leq r$ unknowns per row, $\leq c$ unknowns per column	$\max\{c, r\} + \log m$	$\log m$
	Few Knowns: $\lesssim \sqrt{n}$ knowns per row (hypotheses)	$\log m$	

Table 2: **Summary of Our Results.** Here, "row" and "column" refer to the outcome table. "Apxn. ratio" is the worst-case ratio (over all instances) between the expected cost of our algorithm and the optimal cost achievable by the same type of adaptivity (i.e., non-adaptive or adaptive).

Despite the extensive literature on ODT, little is known about the noisy version from the perspective of approximation factor. There are two main reasons. First, the persistence of noise disables most of the statistical learning tools such as concentration bounds. Secondly, the structure of the optimal solution becomes significantly more complicated under noisy outcomes, posing substantial challenge for the analysis of approximation ratio.



0.09	0.01	0.04	0.34	0.05	0.56	0.24	0.10	0.07	0.14	0.01	0.24	0.05	0.25	0.01	0.30	0.02	0.03	0.18	0.12
0.26	0.17	0.68	0.16	0.25	0.01	0.06	0.25	0.05	0.12	0.23	0.49	0.26	0.01	0.23	0.15	0.01	0.05	0.25	0.15
0.01	0.09	0.47	0.69	0.06	0.66	0.41	0.12	0.15	0.02	0.32	0.26	0.15	0.16	0.36	0.34	0.26	0.01	0.12	0.08
0.01	0.28	0.03	0.46	0.01	0.26	0.19	0.44	0.09	0.01	0.31	0.31	0.22	0.04	0.04	0.01	0.15	0.15	0.44	0.21
0.18	0.03	0.11	0.14	0.07	0.3	0.17	0.05	0.02	0.05	0.26	0.1	0.03	0.45	0.26	0.43	0.07	0.06	0.16	0.12
0.29	0.18	0.2	0.08	0.2	0.69	0.04	0.30	0.06	0.2	0.06	0.15	0.19	0.23	0.35	0.36	0.32	0.09	0.14	0.10
0.01	0.05	0.01	0.02	0.09	0.09	0.21	0.09	0.03	0.09	0.3	0.46	0.19	0.07	0.35	0.06	0.15	0.02	0.24	0.06
0.2	0.45	0.09	0.08	0.26	0.15	0.01	0.17	0.02	0.05	0.02	0.02	0.45	0.12	0.4	0.36	0.01	0.06	0.08	0.01

Figure 1: **Example Mutation Table:** *Each row/column correspond to a cancer type/genetic interval. The entries correspond to the mutation probability of a genetic interval (action) under a cancer type (hypothesis). The actual table entries are often much lower.*

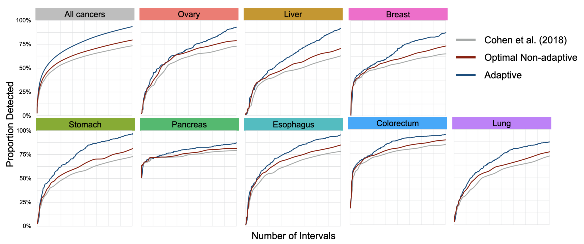


Figure 2: **Empirical Performance.** *Comparison of (non-adaptive) genomic panels from* **?** *with optimal non-adaptive panels and adaptive panels constructed using our partially adaptive greedy algorithm. The detection rate (on the COSMIC dataset) is plotted as a function of panel size.*

Thus motivated, in Jia et al. [2019], we study a fundamental variant of the ODT problem where some test outcomes are noisy, even in the more general case where the noise is *persistent*, i.e., repeating a test gives the same noisy output. Our approximation algorithms provide guarantees that are nearly best possible and hold for the general case of a large number of noisy outcomes per test or per hypothesis where the performance degrades continuously with this number. More generally, we introduce a problem, *Submodular Function Ranking with Noise*, which further generalizes the above.

We design new approximation algorithms for both the non-adaptive setting, where the test sequence must be fixed *a-priori*, and the adaptive setting where the test sequence depends on the outcomes of prior tests. Our new approximation algorithms provide guarantees that are nearly best-possible and work for the general case of a large number of noisy outcomes per test or per hypothesis, with performance degrading smoothly with this number. Moreover, our numerical evaluations show that despite our theoretical logarithmic approximation guarantees, our methods give solutions with cost very close to the information theoretic minimum.

Building upon this work, and inspired by recent advancements of liquid biopsies, in Gan et al. 2021 we study the *active sequential hypothesis testing* (ASHT) problem, which is essentially a variant of the noisy ODT problem with a prescribed noise tolerance level. Specifically, given an *error budget* $\delta > 0$, we aim to identify the true hypothesis with probability at least $1 - \delta$.

Partially Adaptive	**Brute Force**	**LP Heuristic**	**Our Algorithm**
Runtime	$\Omega(\|A\|!)$	LP of size $\Omega(\|A\|\|H\|^2)$ once	$O(\|A\|\|H\|)$ per iter.
Approximation Ratio	1	—	$O(\log(\|H\|))$
Fully Adaptive	**Brute Force**	**?**	**Our Algorithm**
Runtime	$\Omega(\|A\|^{2^{\|H\|}})$	LP of size $\Omega(\|A\|\|H\|)$ per iter.	$O(\|A\|\|H\|)$ per iter.
Approximation Ratio	1	—	$O(\log^2(\|H\|))$

Table 3: **Summary of Theoretical Results:** *We compare the performance of our algorithms with benchmark algorithms, in terms of runtime and approximation guarantees. The bound of our algorithm is stated for fixed separation parameter and error tolerance.*

Motivated by applications in which the number of hypotheses or actions is massive (e.g., genomics-based cancer detection), we propose efficient greedy algorithms and provide the first approximation guarantees for ASHT. Our guarantees are independent of the number of actions and logarithmic in the number of hypotheses. On a real-world DNA mutation data (COSMIC), our algorithms substantially outperform previously proposed heuristic policies; see Figure 2. This work received the 2021 **Pierskalla Best Paper Award** in Healthcare Applications for its application in liquid biopsy, an emerging cancer detection method.

3 Short-Lived High-Volume Bandits

Recommendation tasks can be classified into four categories based on the *lifetime* and *volume* of contents generated. For persistent (long-lived) content, the problem is arguably straightforward: spend a small amount of time collecting sufficient data in the form of user feedback, and then apply suitable offline predictive model, which may range from a basic collaborative filtering algorithms to deep neural networks (DNNs). Orthogonal to content lifetime, when there is a *low volume* of content relative to the number of users, the problem is similarly well-understood: dedicated exploration methods (e.g. A/B testing) are sufficient for finding the right segments of users for which the content is most appealing.

Naturally then, the most challenging settings are where the content to be recommended is *short-lived* and *high-volume*. Such settings arise, for example, in content aggregation platforms (e.g. Apple News) and platforms with content that is entirely user-generated (e.g. TikTok). In these settings, both previous approaches are prone to failure: offline predictive algorithms do not receive enough data on individual content to achieve meaningful accuracy due to the short lifetime, and dedicated exploration methods are ill-suited to high volume.

Our collaboration with *Glance* (see Figures 3 and 4), a subsidiary of Inmobi - the first unicorn in India [Kadakia, 2023] - faces exactly this challenge. As a leading lockscreen content platform, their marketing team curates a large number of *content cards* per hour, which will be sent to the users' phone. Most content cards have a short lifetime due to their transient nature, making the problem more challenging.

Figure 3: **Sample Glance Cards:** *A Glance card typically consists of an image or graphic, a brief description, and a call-to-action (CTA). The contents vary from news, entertainment and advertisement.*

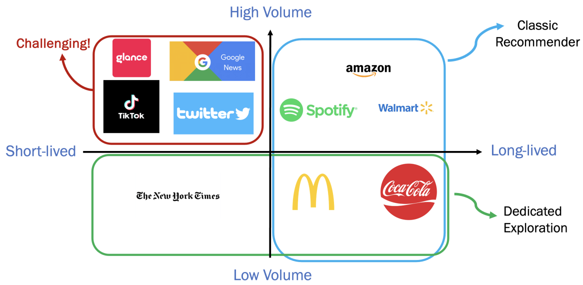


Figure 4: **Lifetime and Volume.** *Scenarios where items are long-lived (blue) or arrive in low volume (green) are relatively easy. This work focuses on the challenging scenario (red) where items are short-lived and arrive in high volume.*

We tackle this challenge systematically in Jia et al. [2023b] by introducing a multiple-play Bayesian bandit problem that encapsulates the key features of this challenge. In each round, $O(n^\rho)$ new arms arrive. Each arm is available for a short lifetime w and has an unknown reward rate, which is drawn from a distribution with a density bounded above and below (away from 0), and may vary over time. The learner selects a multi-set of n arms in each round and receives observable rewards. We present a policy whose *loss* (due to not knowing the reward rates) decreases in w, eventually converging to a lower bound as $n \to \infty$ for any fixed $\rho > 0$.

Notably, we also validated the effectiveness of our policy through a large-scale field experiment on *Glance*. We implemented our policy with some practical adjustments on around 1% of their total traffic over 14 days. This involved around 510,000 users and 18 million impressions. Through a comprehensive statistical analysis, we found that our policy outperformed their DNN-based recommender by 4.32% in total duration and 7.48% in total click-throughs.

References

Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation (TEAC)*, 3(1):1–26, 2015.

Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

Ningyuan Chen. Multi-armed bandit requiring monotone arm sequences. *Advances in Neural Information Processing Systems*, 34:16093–16103, 2021.

Kyra Gan, Su Jia, Andrew Li, and Sridhar R Tayur. Toward a liquid biopsy: Greedy approximation algorithmsfor active sequential hypothesis testing. *Available at SSRN*, 2021.

Su Jia, Viswanath Nagarajan, Fatemeh Navidi, and R. Ravi. Optimal decision tree with noisy outcomes. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3298–3308, 2019.

Su Jia, Andrew Li, and R Ravi. Markdown pricing under unknown demand. *Available at SSRN 3861379*, 2021.

Su Jia, Andrew Li, and R Ravi. Markdown pricing under an unknown parametric demand model. *arXiv preprint arXiv:2312.15286*, 2023a.

Su Jia, Nishant Oli, Ian Anderson, Paul Duff, Andrew A Li, and Ramamoorthi Ravi. Short-lived high-volume bandits. In *International Conference on Machine Learning*, pages 14902–14929. PMLR, 2023b.

Pankti Mehta Kadakia. Inmobi's glance launches in japan, aims for 40 percent of android market. *Forbes India*, 2023. URL `https://www.forbesindia.com/article/take-one-big-story-of-the-day/inmobis-glance-launches-in-japan-aims-for-40-percent-of-android-market/87801/1`.

Robert D Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2005.

Michael Luca and Oren Reshef. The effect of price on firm reputation. *Management Science*, 2021.