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 2, 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 exploring the 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 and the price), they may simply implement the revenue-maximizing policy. However, the demand function may not be available in practice, so they need to experiment with different prices to learn the demand function, and make pricing decisions based on the observed data.

The above *learning vs. earning*, or *exploration vs. exploitation* trade-off can be captured by the *Multi-Armed Bandits* (MAB) framework, which has attracted significant attention from a range of communities. 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 by addressing three challenging aspects:
(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 stem from fundamental areas at the interface of operations and marketing, including pricing, survey design, and content recommendation. We offer theoretical insights with performance guarantees and empirical evidence demonstrating why our methodologies outperform existing approaches.

1 Markdown Pricing Under Unknown Demand

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

This motivates us to consider a dynamic **markdown** pricing problem, where the price sequence is required to be non-increasing. While (a) markdown pricing under *known* demand and (b) *unconstrained* pricing (i.e., where the price is allowed to increase and decrease) under unknown demand have been extensively studied, little is known for markdown pricing under an unknown demand function. In particular, some basic questions remain open:

What is the optimal regret for markdown pricing under unknown demand?
And are they higher than the bounds for unconstrained (i.e., 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 $r(p)$ is the product of the price p and the demand function $d(p)$. The unknown demand function comes from a given family, which may be either nonparametric [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]		yes	yes	$\tilde{\Theta}(n^{k/k+1})$
	Broder and Rusmevichientong [2012]	0-crossing	yes	no	$\Theta(\log n)$
	Jia et al. [2023a]		yes	yes	$\Theta(\log^2 n)$

Table 1: **Minimax Regret, With and Without Monotonicity.** Our results (**blue**) are **optimal** up to logarithmic (or constant) terms. Here, r'' is the second derivative of the revenue function r. Non-parametric results still hold if the initial inventory m is finite, with n replaced by $\max\{m, n\}$.

We provide a **complete** settlement by presenting novel learning policies with minimax optimal regret, summarized in Table 1. Notably, in the parametric setting, 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 in the family do not intersect. We highlight that most of these bounds have higher asymptotic orders than their unconstrained counterparts, except for the 1-crossing family (e.g., linear demand function).

2 Optimal Decision Tree with Noisy Outcomes

Combinatorial structures in the problem instance, such as those enabling binary search, can sometimes be leveraged to speed up learning. This is encapsulated by the *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

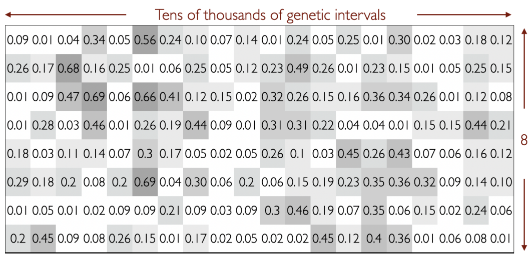
Adaptivity	Assumption	Apxn. ratio	APX-hardness
non-adaptive	none	$\log m$	$\log m$
adaptive	**Few Unknowns:** $\leq r$ unknowns per row, $\leq c$ unknowns per column	$\max\{c, r\} + \log m$	$\log m$
	Few Knowns: $\lesssim \sqrt{n}$ knowns per row (hypotheses)	$\log m$	

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).

of test and hypothesis, we aim to find a low-cost testing procedure (*decision tree*) that identifies the true hypothesis, with an acceptable misidentification probability.

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. Despite the extensive literature on ODT, little is known about the noisy version from the perspective of approximation algorithms. There are two main reasons: First, the persistence of noise disables many 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.

Thus motivated, in Jia et al. [2019], we study a fundamental variant of the ODT problem where some test outcomes are uncertain (stochastically or adversarially), even in the persistent-noise setting. We design new approximation algorithms for both the non-adaptive setting, where the test sequence must be fixed in advance, and the adaptive setting where the test sequence depends on the outcomes of prior tests. Our new approximation algorithms offer nearly optimal guarantees and handle cases with many noisy outcomes per test or hypothesis, with performance gradually decreasing as uncertainty increases. Moreover, our numerical evaluations show that our methods give solutions with cost very close to the information theoretic optimum, despite our theoretical logarithmic approximation guarantees.



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.*

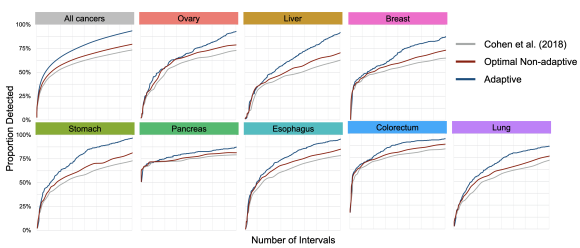


Figure 2: **Empirical Performance.** *Comparison of (non-adaptive) genomic panels from Cohen et al. [2018] with optimal non-adaptive and adaptive panels constructed using our greedy algorithms. The detection rate on the COSMIC data set is plotted as a function of panel size.*

Inspired by recent advancements of liquid biopsies, in Gan et al. 2021 we study the *Active Sequential Hypothesis Testing* (ASHT) problem. Essentially, this a variant of the noisy ODT problem with a prescribed misidentification probability. Specifically, given $\delta > 0$, we aim to identify the true hypothesis with probability at least $1 - \delta$.

Partially Adaptive	**Brute Force**	**LP Heuristic**	**Our Algorithm**
Runtime	$\Omega(\lvert A\rvert!)$	LP of size $\Omega(\lvert A\rvert \lvert H\rvert^2)$ once	$O(\lvert A\rvert \lvert H\rvert)$ per iter.
Approximation Ratio	1	—	$O(\log(\lvert H\rvert))$
Fully Adaptive	**Brute Force**	Naghshvar and Javidi [2013]	**Our Algorithm**
Runtime	$\Omega(\lvert A\rvert^{2^{\lvert H\rvert}})$	LP of size $\Omega(\lvert A\rvert \lvert H\rvert)$ per iter.	$O(\lvert A\rvert \lvert H\rvert)$ per iter.
Approximation Ratio	1	—	$O(\log^2(\lvert H\rvert))$

Table 3: **Summary of Results:** *We compare the performance of our algorithms with benchmark algorithms, in terms of runtime and approximation guarantees. Our bounds are stated for a fixed error tolerance.*

Motivated by applications where the numbers of hypotheses and actions are massive (e.g., genomics-based cancer detection), we propose efficient greedy algorithms and provide the first approximation guarantees for ASHT under two types of adaptivity (partially adaptive and fully adaptive); see Table 3. Notably, our guarantees are independent of the number of actions and logarithmic in the number of hypotheses. Moreover, on a real-world DNA mutation data (COSMIC), our algorithms substantially outperform previous heuristics; see Figure 2. Our work received the 2021 **Pierskalla Best Paper Award** in Health Applications.

3 Short-Lived High-Volume Bandits

Recommendation tasks can be classified into four categories based on the *lifetime* and *volume* of contents generated, as shown in Figure 4. For persistent (long-lived) content, the problem is arguably straightforward: Collect sufficient amount of data, and apply a suitable offline predictive model, e.g., collaborative filtering or deep neural network (DNN). Orthogonal to the content lifetime, when there is a *low volume* of content relative to the number of users, the problem is also well understood: Dedicated exploration methods (e.g., basic A/B testing) are sufficient for identifying the most appealing content for each segment of users.

The most challenging settings then 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 lack sufficient data on individual content to achieve meaningful accuracy due to its short lifetime, and dedicated exploration methods are ill-suited since only a limited number of samples are collected for each piece of content on average.

Our collaborator, *Glance* - a subsidiary of the first unicorn *Inmobi* 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 (see Figures 3 and 4), 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.

We tackle this problem systematically in Jia et al. [2023b] by introducing a multiple-play Bayesian

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.*

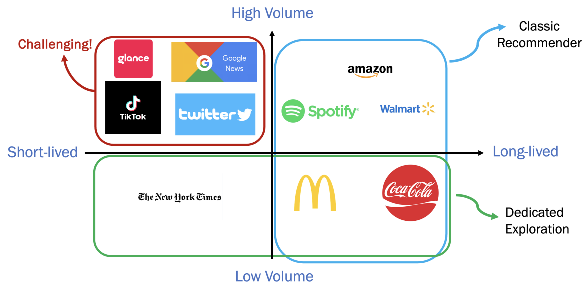


Figure 4: **Lifetime and Volume.** *Scenarios where items are long-lived (blue) or arrive in low volume (green) are relatively easy. This work focuses on the challenging scenario (red) where items are short-lived and arrive in high volume.*

bandit problem that encapsulates the key features described above. In each round, $O(n^\rho)$ new arms (content cards) arrive. Each arm is available for a short lifetime w (after which it expires), and has an unknown reward rate drawn from a distribution $\mathcal{D}$ that may vary over time. The learner selects a multi-set of n arms in each round and receives observable rewards at the end of the round. 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 $\rho > 0$.

Finally, we validated the effectiveness of our policy through a large-scale field experiment on Glance's real system, conducted on approximately 1% of their total traffic over 14 days in 2021. This involved approximately $510,000$ users and 18 million impressions. With certain practical adjustments, 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.

Joshua D Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378):926–930, 2018.

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.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. 2013.