Learning to Select and Rank from Choice-Based Feedback: A Simple Nested Approach[*]

Junwen Yang
National University of Singapore
Singapore 119245
junwen_yang@u.nus.edu

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

Abstract

We study a ranking and selection problem of learning from choice-based feedback with dynamic assortments. In this problem, a company sequentially displays a set of items to a population of customers and collects their choices as feedback. The only information available about the underlying choice model is that the choice probabilities are consistent with some unknown true strict ranking over the items. The objective is to identify, with the fewest samples, the most preferred item or the full ranking over the items at a high confidence level. We present novel and simple algorithms for both learning goals. In the first subproblem regarding best-item identification, we introduce an elimination-based algorithm, NESTED ELIMINATION (NE). In the more complex subproblem regarding full-ranking identification, we generalize NE and propose a divide-and-conquer algorithm, NESTED PARTITION (NP). We provide strong characterizations of both algorithms through instance-specific and non-asymptotic bounds on the sample complexity. This is accomplished using an analytical framework that characterizes the system dynamics through analyzing a sequence of multi-dimensional random walks. We also establish a connection between our nested approach and the information-theoretic lower bounds. We thus show that NE is worst-case asymptotically optimal, and NP is optimal up to a constant factor. Finally, numerical experiments from both synthetic and real data corroborate our theoretical findings.

Understanding customer preferences is fundamental to decision-making across domains such as marketing, e-commerce, and recommendation systems. Advances in internet and computing technologies have significantly enhanced the sophistication of preference learning systems, enabling them to operate in real time, adapt dynamically, deliver personalized results, and scale efficiently. These developments have unlocked novel applications. For example, a business model innovation in e-commerce is *crowdvoting*, where companies systematically collect consumer feedback on new product prototypes to decide which to bring to market (see 14, 17, 1). More broadly, digital surveys have become increasingly prevalent, allowing firms to better understand customer preferences. These trends underscore the importance of designing efficient preference learning systems. For instance, well-designed feedback mechanisms can help crowdvoting avoid delays in product introduction and reduce the risk of launching poorly received products. In digital surveys, efficient data collection is crucial, as participant compensation makes sample inefficiency financially burdensome.

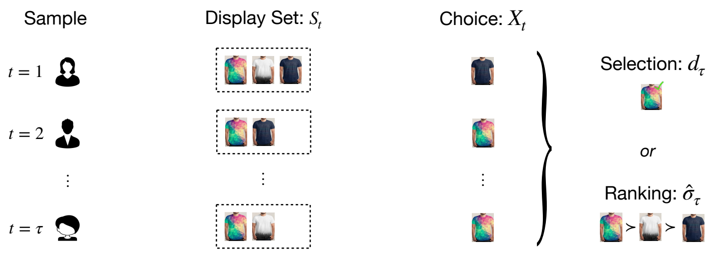
Motivated by those preference learning applications, we investigate a class of ranking-and-selection problems from a specific feedback structure, which we refer to as *choice-based feedback*. To illustrate, consider a company seeking to understand customer preferences among a set of items (e.g.,

[*] A preliminary version of this work appeared in the *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Current version: https://arxiv.org/abs/2307.09295.

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.

product prototypes for commercialization). The company may pursue one of two objectives: identifying the best item or ranking the entire set of items. To achieve these goals, the company can present subsets of items to customers, asking them to select their favorite within each set. The company can dynamically adjust these display sets based on previous feedback. The central challenge lies in designing these display sets to make the learning process *efficient* – minimizing the cost of feedback collection while ensuring high accuracy in the final outcomes; see Figure 1 for an illustration and the full paper for a formal description.

Figure 1: A Visualization of The Learning to Select (and Rank) Problem.



Decisions: Display policy + stopping rule + final selection/ranking

Goal: Use the fewest samples to $\begin{cases} \text{select the top-ranked item} \\ \text{ranked the items} \end{cases}$ with high probability.

Choice-based feedback offers both opportunities and challenges. On the one hand, choices and comparisons provide a natural and intuitive form of feedback. Its advantages over alternative formats such as ratings or scores are discussed across various disciplines such as opinion research (15), psychology (9) and computer science (19). On the other hand, the combinatorial nature of display sets (also known as assortments) introduces significant complexity, especially when combined with dynamic learning aspects. Systematic studies of this problem remain relatively nascent.

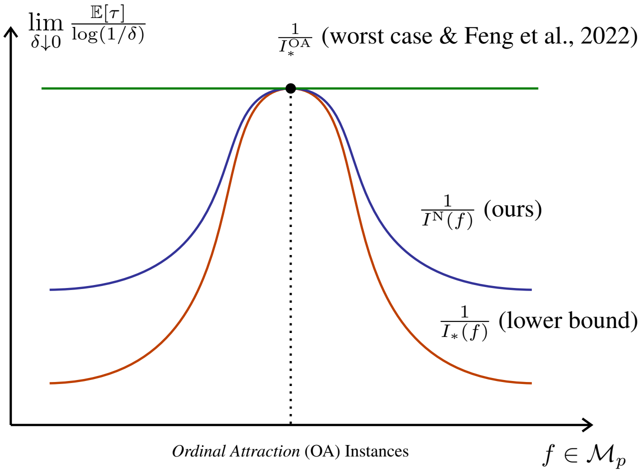
In this regard, our paper contributes to an emerging literature that brings machine learning and operations research tools to this type of problem (see, e.g., 18, 3, 6, 1, 7). Among these, the work by [6] is most closely related to ours. They introduced a relatively general framework for modeling customer preferences through discrete choice probabilities. Instead of parametric choice models such as Multinomial Logit (MNL), they only imposed certain consistency and separability conditions on the choice probabilities, namely, a more preferred item is chosen with strictly higher probabilities. Under this modeling framework, they studied a best-item identification problem under the fixed-confidence setting, i.e., aiming to minimize the feedback required to guarantee a desired level of confidence. Leveraging an information-theoretic measure dating back to [4], they proposed a randomized policy, MYOPIC TRACKING POLICY (MTP), and showed that it is worst-case asymptotically optimal. Their work also highlighted a useful trade-off in this problem: larger display sets increase coverage by comparing more items simultaneously but may reduce the precision of individual comparisons. Conversely, smaller sets (e.g., pairwise displays) enhance precision but limit coverage. Hence an optimal procedure should harvest the benefit of both.

Summary of Main Results and Contributions

The first part of our paper revisits the best-item identification problem, also referred to as "learning-to-select." It is motivated by several notable limitations of the MTP policy by [6]. One pressing issue is that it requires repeatedly solving combinatorial optimization problems throughout the time horizon, which restricts the scalability of the algorithm. Furthermore, the theoretical guarantee of MTP has two important drawbacks: (i) it focuses on the hardest-to-learn instances with limited insights for general cases, and (ii) it allows a residual term on the order of $o(\log(1/\delta))$, where δ is the target error probability. These two limitations imply that the guarantees of MTP may be weak for general instances and when the target error probability is only moderately small.

We propose a surprisingly simple algorithm, NESTED ELIMINATION (NE). It significantly improves upon earlier approaches by (i) being computationally simpler and (ii) offering stronger theoretical guarantees. Our main contributions are as follows:

Figure 2: A Conceptual Illustration of Theoretical Contributions of NE.



Notes: The horizontal axis represents different preference instances f; the vertical axis represents the asymptotic sample complexity.

1. *Simpler Implementation.* NE employs a "nested" structure, shrinking display sets on a path-wise basis. This is combined with a carefully designed (but easy-to-implement) sequence of hitting times that determine when and how suboptimal items are eliminated. By avoiding the need to solve combinatorial optimization problems, NE achieves a running time reduction of up to *three orders of magnitude* compared to MTP; see the full paper.

2. *Stronger Theoretical Guarantee.* We provide a thorough theoretical analysis of NE's performance. For *every* preference instance f (not just worst-case one) and every error tolerance δ, we derive a non-asymptotic and instance-specific bound on its sample complexity.

   Theorem. *For every confidence level $\delta \in (0,1)$,* NE *identifies the top-ranked item with probability at least $1-\delta$. Furthermore, for every preference instance f, the sample complexity satisfies*

   $$\mathbb{E}[\tau] \le \frac{\log(1/\delta)}{I^{\mathrm{N}}(f)} + C_f,$$

   where $I^{\mathrm{N}}(f)$ is an explicit function of the instance f and C_f is a constant independent of δ.

   This bound universally outperforms that of MTP, where the improvement can be up to the order of $\Omega(\log(1/\delta))$. Furthermore, by comparing with the information-theoretical lower bound, we show that NE achieves higher-order worst-case optimality than MTP, where the "sensitivity" of the optimality criterion sharpens from $O(\log(1/\delta))$ to $O(1)$. (See the full paper for more details.)

In the second part of the paper, we consider the more challenging full-ranking identification problem, which we refer to as "learning-to-rank." We introduce a divide-and-conquer algorithm named NESTED PARTITION (NP); see the full paper. The elimination procedure NP mirrors that of the well-known Quicksort algorithm [10] and similarly recursively partitions the active set into two parts, where items in one part are deemed superior to those in the other. We theoretically establish NP's sample complexity.

Theorem. *For every confidence level $\delta \in (0,1)$,* NP *identifies the full ranking with probability at least $1-\delta$. Furthermore, for every preference instance f, the sample complexity satisfies*

$$\mathbb{E}[\tau] \le \frac{\log(1/\delta)}{J^{\mathrm{N}}(f)} + C'_f,$$

where $J^N(f)$ is an explicit function of the instance f and C'_f is a constant independent of δ

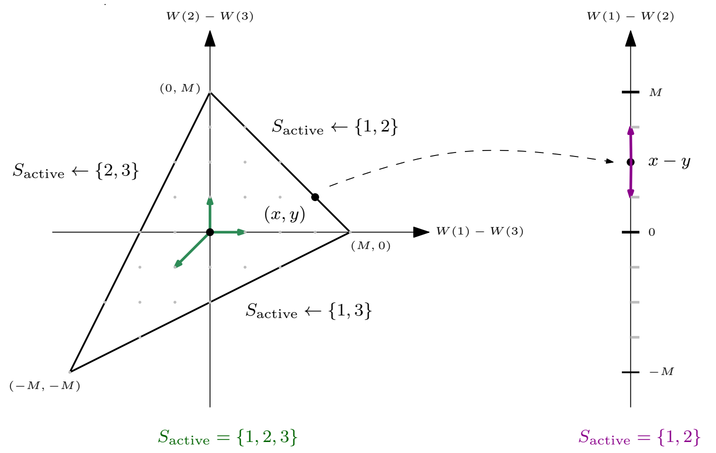
By comparing with the information-theoretic lower bound for the full-ranking identification problem, we show that NP attains (nearly) worst-case asymptotic optimality.

Methodological Innovations. We also find it helpful to briefly outline the main technical challenges and describe how we address them methodologically. Let us begin with the challenges:

1. While the nested approach is intuitive and offers simple structures, it is not immediately clear *a priori* whether such nested structures are optimal. and if so, in what precise sense.
2. Even within the nested framework, many important design choices remain unresolved. For instance, when should an item be eliminated? And if upon elimination, which item should be removed? These questions ultimately reduce to a sequence of stopping problems, the solutions to which are also not evident *a priori*.
3. Although NE and NP are apparently simple to implement, they are nontrivial to *analyze*. The history-dependent elimination rules and the requirement to "transfer" information across different assortments complicate efforts to decouple the dynamics across items or stages. This interdependence renders the overall system dynamics difficult to characterize analytically.

We address the first challenge by drawing a connection between our nested approach and the nested structure of the optimal solution to a max-min problem involving a [4]-type information-theoretic criterion. In this light, our strategy of pathwise shrinking of display sets is not ad hoc. Rather, the optimal allocation among display sets exhibits a "natural" nested structure. To tackle the second challenge, we relate our elimination rules to a specific class of sequential probability ratio tests (SPRTs), with the appropriate hypotheses to test and the right choice model classes.

Figure 3: A Visualization of the Nested Random Walk Dynamics Under NE.


$W(2) - W(3)$
$W(1) - W(2)$
$(0, M)$
$S_{\text{active}} \leftarrow \{1, 2\}$
M
$S_{\text{active}} \leftarrow \{2, 3\}$
$x - y$
(x, y)
$W(1) - W(3)$
0
$(M, 0)$
$S_{\text{active}} \leftarrow \{1, 3\}$
$-M$
$(-M, -M)$
$S_{\text{active}} = \{1, 2, 3\}$
$S_{\text{active}} = \{1, 2\}$

Notes: The NE algorithm maintains an *active* set, denoted by S_{active}, which initially contains all items and shrinks over time. At each time step t, NE displays S_{active} to the next customer and records the observed choice $X_t \in S_{\text{active}}$. The algorithm relies on a simple sufficient statisticthe *voting scores* $\{W(i)\}$ for each item $i \in [K]$which count the number of times each item is chosen.

In the illustrated case with $K = 3$, the first stage starts with the active set $[3] = \{1, 2, 3\}$. The systems dynamics are visualized by projecting $\{W(i)\}$ onto the two-dimensional space spanned by $(W(1) - W(3),\ W(2) - W(3))$. This yields a two-dimensional random walk originating at the origin. The walk evolves via i.i.d. increments of $(0, 1)$, $(1, 0)$, or $(-1, -1)$, occurring with respective probabilities $f(1 \mid [3])$, $f(2 \mid [3])$, and $f(3 \mid [3])$. The first stage ends when the random walk hits the boundary of a triangle with vertices $(0, M)$, $(M, 0)$, and $(-M, -M)$, each face corresponding to the elimination of one item. In the example path, item 3 is eliminated, reducing the active set to $[2] = \{1, 2\}$.

In the second stage, the state is further projected onto the one-dimensional space defined by $W(1) - W(2)$, yielding a one-dimensional random walk. This walk starts from the endpoint inherited from the first stage and evolves by increments of $+1$ or -1, with probabilities $f(1 \mid [2])$ and $f(2 \mid [2])$, respectively. The stage concludes when the walk reaches either M or $-M$, signifying the selection of item 1 or 2, respectively.

To address the third challenge, we represent the system dynamics by a sequence of multi-dimensional random walks, with each later-stage walk initialized at the terminal state of the previous stage. Figure 3 illustrates this for NE; a corresponding illustration for NP is provided in the full paper. This reformulation enables us to reduce the analysis to characterizing the hitting times and hitting distributions of these random walks at each stage. Leveraging tools such as martingale theory, we are able to derive sharp performance bounds, leading to residual terms on the order of $O(1)$, as opposed to the more conventional order of $o(1/\log(\delta))$.

Both our algorithmic design and analytical techniques depart significantly from classical successive elimination-based methods for multi-armed bandits or the widely adopted track-and-plug-in strategies in pure-exploration problems. As such, we believe our approach holds independent value and serves as a useful ground for the development of online learning algorithms for other purposes.

References

[1] Victor F Araman and René A Caldentey. Diffusion approximations for a class of sequential experimentation problems. *Management Science*, 68(8):5958–5979, 2022.

[2] Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 268–276, 2008.

[3] Xi Chen, Yuanzhi Li, and Jieming Mao. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2504–2522. SIAM, 2018.

[4] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

[5] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(6), 2006.

[6] Yifan Feng, René Caldentey, and Christopher Thomas Ryan. Robust learning of consumer preferences. *Operations Research*, 70(2):918–962, 2022.

[7] Yifan Feng and Yuxuan Tang. A mallows-type model for preference learning from (ranked) choices. *Available at SSRN 4539900*, 2023.

[8] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[9] Richard D Goffin and James M Olson. Is it all relative? comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, 6(1):48–60, 2011.

[10] Charles AR Hoare. Quicksort. *The computer journal*, 5(1):10–16, 1962.

[11] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30, 1963.

[12] Shivaram Kalyanakrishnan and Peter Stone. Efficient selection of multiple bandit arms: Theory and practice. In *ICML*, 2010.

[13] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

[14] Andrew King and R. Karim Lakhani. Using open innovation to identify the best ideas. *MIT Sloan Manag. Rev.*, Sep 2013.

[15] Jon A Krosnick and Duane F Alwin. A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, 52(4):526–538, 1988.

[16] Mo Liu, Junyu Cao, and Zuo-Jun Max Shen. Value of one data point: Active label acquisition in assortment optimization. *Available at SSRN 4487888*, 2023.

[17] Simone Marinesi and Karan Girotra. Information acquisition through customer voting systems. 2013.

[18] Sahand Negahban, Sewoong Oh, Kiran K. Thekumparampil, and Jiaming Xu. Learning from comparisons and choices. *Journal of Machine Learning Research*, 19(40):1–95, 2018.

[19] Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. When is it better to compare than to score? *arXiv preprint arXiv:1406.6618*, 2014.

[20] Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 109–117. PMLR, 2013.