Submitted to *Operations Research*
manuscript

Robust Learning of Consumer Preferences

Yifan Feng[1] René Caldentey[1] Christopher Thomas Ryan[2]

[1]University of Chicago, Booth School of Business, email: {yifan.feng,rene.caldentey}@chicagobooth.edu

[2] University of British Columbia, Sauder School of Business, email: chris.ryan@sauder.ubc.ca

This paper studies a class of ranking and selection problems faced by a company that wants to identify the most preferred product out of a finite set of alternatives when consumer preferences are *a priori* unknown. The only information available is that consumer preferences satisfy two key properties: (*i*) they are consistent with some unknown true ranking of the alternatives and (*ii*) they are strict, namely, no two products are equally preferred. To learn the unknown ranking, the company is able to sample consumer preferences by sequentially showing different subsets of products to different consumers and asking them to report their top preference within the displayed set. The objective of the company is to design a display policy that minimizes the expected number of samples needed to identify the top-ranked product with high probability. We prove an instance-specific lower bound on the sample complexity of any policy that identifies the top-ranked product within a given (probabilistic) confidence. We also propose a robust formulation of the company's problem and derive a sampling policy (Myopic Tracking Policy), which is both worst-case asymptotically optimal and intuitive to implement. Roughly speaking, the Myopic Tracking Policy randomly alternates between two extreme types of displaying strategies: (i) *full display* that shows a consumer the entire menu so as to learn something about every product and (ii) *pair display* that shows a consumer only two products so as to maximize the informativeness of the choice made by the consumer. To assess the performance of our proposed Myopic Tracking Policy, we conduct a comprehensive set of computational studies and compare it to alternative methods in the literature.

Key words: sequential learning, maximum selection, best arm identification, dynamic assortments, preference learning

1. Introduction

Problem Overview. A company wants to identify the 'best' version of a product to commercialize in the marketplace from a menu $[K] = \{1, 2, \ldots, K\}$ of alternative versions. The company does not know consumer preferences over these versions and implements a consumer feedback system to collect information. Specifically, the platform is able to display different subsets of versions to different consumers, who then give feedback in the form of a *vote* for their most preferred version within the subset they see. In addition, the system must decide when to stop the feedback process and make a recommendation on which version to commercialize.

There are many possible feedback mechanisms that a company can use. Traditional examples include “taste tests”, focus groups, or surveys of potential consumers. With the advent of the Internet, methods of feedback have become more sophisticated. One trend in online commerce is the use of *crowdvoting* platforms to collect consumer feedback about possible new products or other business innovations. For example, Chicago-based fashion company Threadless uses a crowdvoting platform to feature T-shirt designs from freelance designers on a weekly basis and solicits consumer opinions online for preferred designs. Threadless uses this consumer feedback to narrow down the number of designs that are sent to production (Brabham 2010, King and Lakhani 2013).[1]

The task of efficiently managing the feedback platform –*i.e.*, balancing the quality and speed of the learning process– is complex, particularly when (i) the number of alternative versions is large and (ii) making inferences about consumer preferences from votes is limited. In such cases, the company needs to judiciously and dynamically choose which subset of versions to display to each arriving consumer, with the objective of maximizing the *amount of information* generated by each consumer choice. The optimal choice of display sets is contingent on the history of votes observed over time. In terms of the length of the feedback process, the company would like to minimize the time required to make a final recommendation, but without jumping too quickly to a recommendation. Commercializing the ‘*wrong*’ version can be costly (Schneider and Hall 2011).

In this paper, we propose a methodology to (i) dynamically choose which *display set* to show each arriving consumer, (ii) decide when to stop the feedback process, and (iii) select the version to commercialize. This methodology minimizes the amount of feedback needed to achieve a fixed probabilistic guarantee of choosing the best version. Our methodology applies to a general class of choice models where consumer choice probabilities are determined by a fixed (unknown to the company) ranking over the set of versions that represent consumer preferences. We provide a detailed mathematical description of this choice model in Section 4.

To get some intuition for the trade-offs involved, and how our methodology balances them, let us discuss two extreme display strategies. On one extreme, the company can use a *full-display* policy where all versions are displayed to every consumer. This allows the company to learn something about each consumer’s preference over every version. In this regard, a full-display policy maximizes the *coverage* achieved by a display policy. However, under reasonable assumptions on the underlying choice model, the probability that a consumer votes for the best version within a given display set decreases in the cardinality of the set. For instance, consumers can become overwhelmed if many

[1] Other examples include: (i) Amazon, which leverages reader nominations to e-publish books (the *Kindle Scout* platform, Pee 2016); (ii) LEGO, which uses crowdvoting to generate and pick new designs of toy sets (the *LEGO Ideas* platform, Lego 2018); and (iii) Betabrand, which uses both crowdfunding and crowdvoting to solicit design ideas and converts selected designs into real products (Betabrand 2018).

alternatives are displayed, making it harder for them to identify their true most-preferred version. Hence, larger display sets may provide less accurate information than smaller ones. To maximize the *accuracy* of the inference made on each consumer choice, the company can use a *pair-display* policy where only two versions are shown to each consumer. Of course, the choice of which pair of versions to display to each consumer should depend on the history of the feedback process. This choice makes implementing an optimal pair-display strategy substantially more complex than a full-display strategy, which displays the same set of versions to all consumers.

In general, neither the full-display nor the pair-display strategies are optimal. This is as expected, the one optimizes for coverage (at the cost of accuracy) while the other optimizes for accuracy (at the cost of coverage). An optimal strategy must strike the right balance between coverage and accuracy. The goal of this paper is to shed some light on this trade-off.

Summary of Methodology and Results. The development of rigorous methods to learn consumer preferences has been the focus of much research in computer science, economics, marketing, and operations (see Section 2 for a review of related literature). A dominant approach to tackle this problem is to impose a parametric structure on the underlying choice model governing consumer behavior. For instance, Luce-type models –and in particular the Multinomial Choice Model (MNL)– are regularly used (e.g., Sauré and Zeevi 2013a, Chen et al. 2018). This parametric approach, however, puts us in an uncomfortable predicament. In order to learn unknown consumer preferences, we must assume that we have a fair amount of knowledge about the parametric structure of those preferences.[2]

With the aforementioned predicament in mind, we develop an efficient active learning algorithm to learn these preferences '*on the fly*' while imposing minimal parametric assumptions. We tackle this challenge using a worst-case formulation of the problem that relies on a mild separability condition. This worst-case analysis singles out a specific choice model we call the *Ordinal Attraction* (OA) model. The OA model is, in some sense, the "noisiest" (and hence, the hardest) consumer choice model to learn among those satisfying our separability condition. Fortunately, as in robust optimization, where the worst-case distribution often has a tractable structure, the OA model has many attractive analytical features that we exploit to propose a robust and computationally efficient display policy.

Our proposed strategy – the Myopic Tracking Policy in Algorithm 1– judiciously balances the coverage/accuracy trade-off and satisfies two key optimality properties: (i) it chooses the best

[2] See Heckel et al. (2019) for a discussion on the limited benefits of parametric models in the context of rankings from pairwise comparisons.

version *with high probability* and (ii) it *asymptotically* minimizes the amount of consumer feedback needed to make a final recommendation. To be precise, for any small $\delta > 0$, we first derive a lower bound on the number of votes needed by any display strategy to achieve a $1 - \delta$ probability of selecting the top-ranked version (Theorem 1). We then show that as $\delta \downarrow 0$, the Myopic Tracking policy needs an expected number of votes that matches this lower bound (Theorem 6) with up to δ error probability (Theorem 7).

To get a panoramic view of our approach and the challenges that we need to address, we summarize its main characteristics. Loosely speaking, at every consumer arrival, the Myopic Tracking policy goes through the following sequence of steps:

Myopic Tracking policy: Schematic Description

Step 1: Using the available history of consumer votes, identify the 'best' preference ranking over all versions.

Step 2: Using the identified preference in Step 1, and the available data, update a stopping criteria. If the criterion is met, stop the feedback process and select the top-ranked version according to the current ranking. Otherwise, go to Step 3.

Step 3: Randomly (according to a pre-specified distribution) select a display set to show the next consumer. Record the vote outcome and go to Step 1.

In Step 1, the algorithm computes a ranking of the versions that best reflects consumer preferences given the feedback history. There are two issues that need to be handled in this step. First, we need to decide how to identify the 'best ranking' in every iteration. Second, we need to be able to compute this ranking efficiently. Our proposed methodology relies on a Maximum Likelihood Estimation (MLE) criteria to select the best ranking at every iteration.

The stopping criterion in Step 2 is of a threshold-hitting type *à la* Chernoff (1959). Specifically, the Myopic Tracking policy keeps track of a stochastic process that measures the discrepancy between the preference ranking identified in Step 1 and the available data. The algorithm stops as soon as this stochastic process hits a fixed lower bound whose value is appropriately calibrated to ensure that the algorithm will select the best version with probability at least $1 - \delta$. The evolution of the underlying stochastic process is related to the MLE computations used in Step 1.

Turning to Step 3, our proposed policy randomly selects the display set shown to each arriving consumer. In particular, the probability distribution that is used to randomize over the display sets is constant –invariant to the feedback history– up to the permutation of the versions induced by the preference ranking identified in Step 1. It follows that this randomization distribution can be computed offline before the actual implementation of the feedback system. Furthermore, under the OA model, these randomization probabilities depend on the relative ranking of each version within the given display set. Through a closed-form characterization (Theorem 4), we find that the Myopic Tracking Policy restricts this randomization to a small subset of nested display sets, i.e., those including the top n most preferred versions according to the preference ranking identified in Step 1.[3] This means that there are only $K-1$ nested sets that are being considered for display, a much smaller number than the $2^K - K - 1$ possible display subsets of $[K]$.[4] This key property of the Myopic Tracking policy dramatically reduces the complexity of its implementation.

Finally, although we derive the OA model as a worst-case consumer preference, for the purpose of proposing a robust solution to the top-ranked selection problem, the OA model has a number of alternative interpretations. For instance, it can be viewed as a generalization of the pair-wise comparison model commonly used in the tournament literature. In particular, the MLE problem in Step 1 in our Myopic Tracking Policy is equivalent to the classical dispersion minimization criterion in the sense of Young (1988)(Proposition 3). Thus, computationally, the MLE problem and stopping criterion verification problem can be cast as versions of the *weighted feedback arc set problem on tournaments* that admits an effective integer linear programming algorithm (see Section 7.2). The OA model is also connected to voting theory and, specifically, the Condorcet criterion (see Section 7.3).

2. Related Literature

Methodology: In terms of methodology, our paper builds on the following three areas of research:

1. Sequential Hypothesis Testing: At a high level, we interpret our problem as an active, sequential, and composite multi-hypothesis testing problem where each hypothesis corresponds to one version being the top-ranked version. When the experimenter is a passive observer of data, the generalized sequential likelihood ratio test is known to be asymptotically optimal under various settings (e.g., Wald 1973, Chernoff 1972, Draglia et al. 1999, Li et al. 2014). This provides some support for Steps 1 and 2 in the Myopic Tracking Policy, which essentially implement a generalized sequential likelihood ratio test. On the other hand, the sampling rule in Step 3 is motivated by classical results in active hypothesis testing, in particular, the Max-Min problem studied in Chernoff (1959) (see also Naghshvar et al. 2013).

[3] Specifically, let $\sigma : [K] \to [K]$ be the *preference ranking* (*i.e.*, a permutation of the elements in $[K]$) identified in Step 1. The collection $\{S_n^\sigma\}_{n=2}^K$ of nested display sets associated with ranking σ is such that $S_n^\sigma = \{\sigma^{-1}(1), \sigma^{-1}(2), \ldots, \sigma^{-1}(n)\}$.

[4] The possible display sets are all the subsets of $[K]$ excluding the empty set and singletons, which are obviously never optimal to display.

There are, however, several key distinctions in our model that prevent us from directly applying the results in Chernoff (1959). First, Chernoff (1959) considers a different objective criterion that incorporates a penalty term for selecting the wrong hypothesis (in our case the wrong version), while our problem has an explicit hard constraint that bounds the error probability upon stopping. Second, Chernoff (1959) analysis is restricted to settings in which each hypothesis consists of a finite number of possible states while we allow for an infinite number of states. Moreover, there are certain separability conditions that are imposed among the alternative hypotheses in Chernoff (1959), which our model relaxes. Lastly, our worst-case analysis allows us to completely solve the corresponding Max-Min problem, which is intractable if the experimenter uses the naïve LP formulation of Chernoff's Max-Min problem.

2. Ranking and Selection: Our paper also contributes to the emerging literature on ranking and selection from pairwise to multiwise noisy comparisons (e.g., Braverman and Mossel 2008, Ailon 2012, Braverman and Mossel 2009, Jiang et al. 2011, Wauthier et al. 2013, Shah and Wainwright 2017, Falahatgar et al. 2017, Heckel et al. (2019), among others).

   Among the very few papers that also consider a multi-wise comparison setting, the closest paper to ours is probably Chen et al. (2018) that studies a top-k selection problem under a Multinomial Logit (MNL) model. The class of choice models we consider is, in general, different than the one used in Chen et al. (2018), although some results are comparable. For instance, our methodology improves the lower bound on sample complexity in Chen et al. (2018) (Theorem 1.3) for a fixed success rate to a generic error rate of $\delta \in (0,1)$ (Theorem 1). Our lower bound is also asymptotically tight because we can find an admissible policy to match the lower bound when δ is small.

   We also approach the selection problem from a different angle and so the optimality regimes in Chen et al. (2018) and our paper are different. Chen et al. (2018)'s algorithm (nearly) matches their lower bound when (i) fixing the success probability and (ii) ignoring poly-logarithmic factors of K and the reciprocal of MNL parameter gaps (Theorems 1.2 and 1.3). By comparison, our proposed algorithm (Theorem 6) (i) is asymptotically optimal with respect to the error rate δ, even with the coefficient term matched, and (ii) allows parameters other than δ to grow large, as long as they grow not too fast compared to $1/\delta$.
3. Best Arm Identification: Our solution method is also related to the best arm identification (BAI) literature (e.g., Audibert and Bubeck 2010, Bubeck et al. 2011, Gabillon et al. 2012). In a generic BAI problem, the experimenter tries to distinguish the best arm (i.e., the one with the largest expected reward) using as few samples as possible, where a sample corresponds

to pulling an individual arm and observing a realization of that arm's reward distribution. Our problem could be vaguely cast as a BAI problem by treating each version as an arm and each display set as a "super-arm", i.e., a subset of arms. With this interpretation, the company decision is to select which super-arm to pull at every time epoch. Two key distinctions between BAI and our problem are (i) by pulling a super-arm, the company is able to learn something about every arm included in the super-arm, and (ii) after pulling a super-arm, the observed response is a particular arm rather than a realization of the arm's reward distribution. There is a growing awareness of the relationship between best arm identification and active sequential hypothesis testing (Russo 2016, Garivier and Kaufmann 2016, Kaufmann et al. 2016), especially regarding the importance of the Max-Min problem proposed in Chernoff (1959).

Applications: In terms of applications, our paper is related to two streams of work:

1. Crowdvoting/Wisdom of Crowd: Our paper brings an optimal learning view to crowdvoting or more generally, leveraging the "wisdom of the crowd" to help with product offering decisions (e.g. Raykar et al. (2010), Marinesi and Girotra (2012), Huang et al. (2014), Araman and Caldentey (2016), to name a few). For example, Marinesi and Girotra (2012) study a two-period model where the company uses an online voting platform as an information acquisition mechanism. Araman and Caldentey (2016) decide on the optimal length of the voting period, so as to balance quality of learning and delay cost. Our paper differs from the previous literature, in the sense that we consider consumer choice behavior among many versions, and focus on how to customize each consumer's choice set to maximize the learning speed.
2. Dynamic Assortment Planning with Learning. Our paper is also related to the growing literature on dynamic assortment planning with demand learning (e.g. Caro and Gallien 2007, Rusmevichientong et al. 2010, Ulu et al. 2012, Sauré and Zeevi 2013b, Agrawal et al. 2019, Agrawal et al. 2017, Chen and Wang 2017, among others). The vast majority of this literature formulates the assortment problem as a revenue maximization (or regret minimization) problem and relies on a "learn and earn" approach to solve it (e.g. Rusmevichientong et al. 2010 and Sauré and Zeevi 2013b). A popular strategy is to divide the selling season into two periods: (1) a "pure learning" period in which assortments are offered sequentially to maximize the amount of learning without any revenue consideration, and (2) a "pure earning" period in which a myopic static strategy (based on the knowledge obtained in the pure learning period) is implemented to maximize revenues. As a general rule, the assortments used during the pure learning period have maximal cardinality, typically determined by an exogenously-imposed capacity constraint.

By contrast, our model is solely concerned with maximizing the likelihood of selecting the best version and thus resembles the "pure learning" period previously mentioned. A key insight that emerges from our work is that the exclusive use of maximal cardinality assortments is, in general, suboptimal. The learning process can be accelerated by judiciously balancing the sizes of the display sets over time.

3. Roadmap of Analysis and Results

In this section, we provide a high-level outline of our analysis and main results with the objective of explaining our methodology in simple and intuitive terms. Formal definitions and precise mathematical statements are presented in the sections that follow.

A company wants to identify the product (or version) that is the most attractive to a given population of individuals out of a set of $K \geq 2$ available alternatives. We assume that the preferences of these individuals (consumers) over the different versions are governed by a probabilistic choice model that satisfies two key properties. First, we will assume that there is an underlying (unknown to the company) ranking of the versions and the likelihood that a consumer would select one specific version out of a given subset is consistent (in a probabilistic sense) with this ranking. Second, we will assume that the preferences satisfy a separability condition under which no two versions are equally preferred. We will denote by $\mathcal{M}_p$ the class of consumer preferences that satisfy these requirements, by $f \in \mathcal{M}_p$ a specific consumer preference, and by $f_* \in \mathcal{M}_p$ the true unknown consensus preference of the consumers.

Because preferences are based on product rankings, for each $f \in \mathcal{M}_p$ there exists a unique permutation of the K versions that is consistent with f. We let Σ denote the set of permutations (or rankings) of the set of versions $[K] := \{1, 2, \dots, K\}$ and $\sigma_f \in \Sigma$ denote the unique permutation associated with a preference $f \in \mathcal{M}_p$. It follows that the set of preferences $\mathcal{M}_p$ can be partitioned into a collection $\{\mathcal{M}_p(k) \colon k \in [K]\}$, where each member $\mathcal{M}_p(k)$ of the partition is the set of preferences f for which $\sigma_f(k) = 1$; that is, k is the top-ranked version under σ_f. Thus, the company's problem can be cast as the multi-hypothesis testing problem of deciding which set $\mathcal{M}_p(k)$ contains the true preference f_*. It is worth highlighting that the company is not directly interested in identifying f_*, simply the top-ranked product under it.

To tackle this problem, the company sets a sequential experimentation (or voting) strategy under which individuals from the target population are sequentially exposed to a subset of the versions (a display set) and are asked to select the version they like the most. The display set can vary from individual to individual and the goal of the company is to design a strategy that will identify, as quickly as possible, the true hypothesis with a given probabilistic confidence. Specifically, for a given error tolerance $\delta \in (0,1)$, the company wants to design a display policy that is δ-*accurate*; that is, chooses the true hypothesis with probability at least $1-\delta$.

Our first result (Theorem 1) shows every δ-accurate policy needs at least $\log(1/\delta)/I_*(f_*)$ samples in expectation to learn the true hypothesis, where $I_*(f_*)$ is *Chernoff's information measure* that determines the informativeness of preference f_*. Our second result (Theorem 2) shows — under an additional constraint on the cardinality of $\mathcal{M}_p$ — that there exists a δ-accurate policy that asymptotically, as $\delta \downarrow 0$, uses no more than $\big(\log(1/\delta)+o(\log(1/\delta))\big)/I_*(f_*)$ samples in expectation to learn the true hypothesis. Combined, these two results formalize the intuitive fact that some preferences are easier (or harder) to learn than others and show that Chernoff's information measure $I_*(f)$ quantifies the learning complexity of a given f.

Motivated by the lower and upper bounds on the sample complexity of any δ-accurate policy identified by Theorems 1 and 2, we adopt a worst-case view of the problem and take on the challenge of identifying the set of preferences that minimize Chernoff's information measure $I_*(f)$ over all the preferences in the set $\mathcal{M}_p$, i.e., the preferences that are the hardest to learn. In Theorem 3, we characterize a subset of this class of worst-case preferences that we call the *Ordinal Attraction* (OA) choice model. The OA model has a number of distinguishing properties that we discuss in detail in Section 6. One property worth mentioning here is that, under the OA model, the probability that a given version k is selected within an arbitrary display set S depends exclusively on the relative ranking of k with respect to the other versions in S. Hence, under the OA choice model, consumer preferences have an ordinal structure. It is this property that motivates the name Ordinal Attraction.

We use the worst-case nature of the OA model to formulate a robust version of the company's problem. In particular, we introduce the subset $\mathcal{M}_p^{\mathrm{OA}} \subseteq \mathcal{M}_p$ of OA preferences and look for a δ-accurate policy with minimum expected sample complexity in $\mathcal{M}_p^{\mathrm{OA}}$. To this end, we propose the *Myopic Tracking Policy* (MTP) and show, in Theorem 5, that it is worst-case asymptotically optimal as $\delta \downarrow 0$ within the class of accurate policies not just in $\mathcal{M}_p^{\mathrm{OA}}$ but in the larger set of preferences $\mathcal{M}_p$. Roughly speaking, this means that for any δ-accurate policy π there exists a preference $f_\pi \in \mathcal{M}_p$ where the expected number of samples needed to identify the top-ranked version under f_π using the MTP policy is less than or equal to the expected number of samples needed by policy π. The proof of Theorem 5 is based on Theorems 6 and 7 that show the MTP policy (i) matches the lower bound on the sample complexity in Theorem 1 asymptotically as $\delta \downarrow 0$ and (ii) is δ-accurate. We also show that the Myopic Tracking Policy relies on a simple randomization strategy over a relatively small subset of all possible display sets. Finally, the stopping criteria of the Myopic Tracking Policy is also simple and takes the form of a first hitting time of an appropriate log-likelihood process, very much in the same spirit as Wald's SPRT method.

In Section 8, we use a set of computational experiments to test the performance, in terms of accuracy and sample complexity, of our proposed Myopic Tracking Policy. Using synthetic data, we find that the Myopic Tracking Policy is particularly well suited for an environment where: (i) the number of versions is large; (ii) responses are noisy; and (iii) the tolerance for error probability is low.

4. Model and Problem Formulation

Consumer preferences over the set $[K]$ of available versions are represented by a *consumer choice model* that defines the probability $f(X|S)$ that a consumer will select version $X \in [K]$ when presented with display set $S \subseteq [K]$. The set $\mathcal{S} := \{S \subseteq [K], |S| \geq 2\}$ denotes the collection of all display sets with at least two versions. We refer to f as a consumer preference or simply a *preference*.

We restrict attention to a class of consumers' preferences that satisfy a specific separability condition. Let Σ denote the set of all permutations of the elements in $[K]$, an element $\sigma \in \Sigma$ is called a *ranking*.

Definition 1. (p-Separable Preferences) Let $p \in [0,1)$ be a fixed constant. A preference f belongs to the class $\mathcal{M}_p$ of p-Separable preferences if:

(A-1) `Non-degeneracy:` For any $S \in \mathcal{S}$, $f(X|S) > 0$ if $X \in S$ and $f(X|S) = 0$ otherwise;

(A-2) `Probability Mass Function:` For any $S \in \mathcal{S}$, $\sum_{X \in S} f(X|S) = 1$;

(A-3) `Ranking-based Preference:` There exists a ranking $\sigma_f \in \Sigma$ such that for any $S \in \mathcal{S}$ and any $X, X' \in S$, $f(X'|S) \leq f(X|S)$ if and only if $\sigma_f(X) < \sigma_f(X')$;

(A-4) `(`p`-Separability)` For any $S \in \mathcal{S}$ and any $X, X' \in S$ such that $\sigma_f(X) < \sigma_f(X')$, the preference f satisfies $f(X'|S) \leq p\, f(X|S)$. $\diamond$

Conditions (A-1) and (A-2) are rather intuitive requirements to impose on any probabilistic choice model. Condition (A-3) imposes a minimum level of consistency on the consumers' preferences to ensure that the problem of identifying the top-ranked problem is well defined. Specifically, under this condition, preferences are independent of the display set. Finally, Condition (A-4) is needed for technical reasons to ensure that we can effectively identify the top-ranked version as the number of votes grows large. This condition essentially requires that (i) versions with lower rankings are more likely to be chosen and (ii) consumers are not indifferent between two products in any display set.

The parameter p measures the degree of informativeness of the choice model. For example, in the extreme case of $p = 0$, consumers select the best alternative among the versions in any display set with probability one. In this case, the identification problem is trivial and the company needs to display the entire assortment to a single consumer to identify the best version. If $p=1$, it is possible that the preference is completely uninformative since the uniform preference $f(X|S)=\mathbb{I}\{X\in S\}/|S|$ belongs to the class of 1-Separable choice models.

Remark 1. For any $p\in[0,1)$ and any preference $f\in\mathcal{M}_p$, the ranking σ_f in Condition (A-3) is uniquely defined. Indeed, for every version $k\in[K]$, define $q_k=1-f(k|[K])$ and let $q_{(k)}$ be the k^{th} order statistic of the sequence $(q_1,q_2,\ldots,q_K)$. Then, $\sigma_f(k)=i$ if and only if $q_k=q_{(i)}$. $\diamond$

Remark 2. It is worth noting that the class of consumers' preferences $\mathcal{M}_p$ includes some popular choice models, modulo some separability condition. For example, let $\mathcal{M}^{\mathsf{Luce}}$ be the class of Luce-type preferences for which the attraction scores of the different versions are all different[5]. Then, it is not hard to see that any preference in $\mathcal{M}^{\mathsf{Luce}}$ is p-separable for some value of p. However, for a fixed value of p, the p-separability requirement in Definition 1 limits the subset of Luce-type preferences that are included in $\mathcal{M}_p$. Indeed, for a Luce-type preference with attraction scores $v_1>v_2>\cdots>v_K$, p-separability requires that $v_j\le p^{j-i}\,v_i$ for all $1\le i<j\le K$. Thus, attraction scores need to decay exponentially fast, which can be a restrictive condition in practice if the number of products is large. One possible approach to handle this limitation is to parametrize the value of p in terms of K. For example, if we set $p=1-\frac{1}{K}$ then p-separability would only require that $v_K\le(1-\frac{1}{K})^K\,v_1$ and since $(1-\frac{1}{K})^K$ converges to e^{-1} attraction scores would no longer be required to decay exponentially fast. In section 8, we will investigate numerically the asymptotic performance of our proposed Myopic Tracking policy in asymptotic regimes in which $p\uparrow 1$ and $K\to\infty$, including the case $p=1-\frac{1}{K}$.

The Mallows model is another special class of choice models that is included in $\mathcal{M}_p$[6]. The next result formalizes this observation.

Proposition 1. *Suppose f_M is a consumer preference induced by a Mallows model with concentration parameter θ, then $f_M\in\mathcal{M}_p$ with $p=e^{-\theta}$.*

Despite the versatility of the class $\mathcal{M}_p$ to capture different choice models, such as Luce and Mallows models, it does rely on the assumption that there is a unique ranking that defines the consumers' preference (i.e., condition (A-3) in Definition 1). As a result, settings with different consumer segments that lead to multi-modal consumer preferences cannot be adequately represented by $\mathcal{M}_p$. Of course, there is always the possibility of considering a mixture of p-separable models to capture these situations.$\diamond$

[5] A Luce-type preference f_{L} is defined by a non-negative vector of attraction scores $\{v_1,v_2,\ldots,v_K\}$ (one for each product) and the probability that a consumer selects product i out of an assortment $S\in\mathcal{S}$ is equal to

$$f_{\mathrm{L}}(i|S)=\frac{v_i}{\sum_{j\in S}v_j}.$$

[6] Mallows choice model is a distance-based ranking model, in which consumers' preferences are described by an entire ranking over the products. The probability that an individual consumer has a ranking $\sigma\in\Sigma$ is given by

$$\mathscr{P}(\sigma)\propto e^{-\theta\,d(\sigma,\sigma_0)},$$

where σ_0 is a modal ranking, $d(\cdot,\cdot)$ is the Kendall-Tau distance between rankings and $\theta>0$ is a concentration parameter. Choice probabilities over assortments are derived from the probability distribution $\mathscr{P}$ over rankings (see Désir et al. (2018) for details).

The Company's Decision Problem. Recall that $f_* \in \mathcal{M}_p$ is the *consensus preference*, which is the true (unknown to the company) consumer preferences over the versions in $[K]$. The objective of the company is to design a display strategy to identify $X^* = \sigma_{f_*}^{-1}(1)$, the most preferred version under f_*, as fast as possible. We do not undertake the more ambitious objective of identifying the top k versions (or even the complete ranking σ_{f_*}). Without loss of generality, assume that σ_{f_*} is the identity ranking $\sigma_* := (1, 2, \ldots, K)$.

Consumers arrive sequentially and are indexed by $t = 1, 2, 3, \ldots$ (we use index t to index both time and consumers since only one consumer arrives per time period.) At time t, the company selects a subset $S_t \in \mathcal{S}$ and displays it to the arriving consumer. Consumer t chooses a version $X_t \in S_t$ and the company records his/her choice. We call X_t the "vote" of consumer t. The history of display sets and votes is captured by the filtration $\mathcal{F}_t$, the smallest sigma-algebra generated by $H_t = (S_1, X_1, \ldots, S_t, X_t)$. We also let $\Delta(\mathcal{S})$ denote the set of probability distributions on $\mathcal{S}$.

An admissible policy has three parts:

1. a *display rule*, i.e., a sequence $\{\lambda_t\}_{t=1}^{\infty}$ of probability distributions $\lambda_t \in \Delta(\mathcal{S})$ adapted to the history $H_{t-1} := (S_1, X_1, \ldots, S_{t-1}, X_{t-1})$,
2. a *stopping rule*, i.e., an $\mathcal{F}_t$ stopping time τ for when the feedback process stops, and
3. a *final selection rule*, i.e., $d_\tau \in [K]$ that identifies which version to select in the end.

We let $\pi = (\{\lambda_t\}_{t=1}^{\infty}, \tau, d_\tau)$ denote an admissible policy. A preference $f \in \mathcal{M}_p$ and an admissible policy π induce a probability distribution $\mathbb{P}_f^\pi(\cdot)$ over the history $\{H_t\}$. We also denote by $\mathbb{E}_f^\pi[\cdot]$ the expectation operator under $\mathbb{P}_f^\pi(\cdot)$. With a slight abuse of notation, we may also suppress the superscript, and use the notations $\mathbb{P}_f(\cdot)$ and $\mathbb{E}_f[\cdot]$, when the context is clear.

The company's objective is to implement a policy π that identifies the top-ranked version as quickly as possible. The challenge in determining an optimal policy is the classical sequential learning trade-off between *confidence* (*i.e.*, how sure is the company about selecting the top-ranked version) and *speed* (*i.e.*, how fast can the company reach a final decision). To mathematically formalize this trade-off, we follow the best-arm identification literature (see, e.g., Gabillon et al. 2012) and use a *fixed confidence* approach in which the company's objective is to minimize the (expected) number of votes subject to a hard constraint that upper bounds the probability of selecting the wrong version. To this end, let us define the notion of a δ-accurate policy:

Definition 2. (δ-accurate policy)

(i) For a given $\delta \in [0, 1]$ and a subset of preferences $\mathcal{M} \subseteq \mathcal{M}_p$, we say that an admissible policy π is $\delta(\mathcal{M})$*-accurate* if

$$\mathbb{P}_f^\pi(\tau < \infty) = 1 \qquad \text{and} \qquad \mathbb{P}_f^\pi(d_\tau \neq \sigma_f^{-1}(1)) \leq \delta \qquad \text{for any } f \in \mathcal{M};$$

that is, if the voting process terminates almost surely and the probability of selecting the wrong version (according to any given preference $f \in \mathcal{M}$) is less than or equal to δ.

(ii) We say that an admissible policy π is δ-*accurate* if it is $\delta(\mathcal{M}_p)$-*accurate.*
(iii) Let $\Pi = \{\pi_\delta\}_{\delta\in(0,1]}$ be a class of admissible policies parameterized by δ. We say that Π is *accurate* if π_δ is δ-accurate for all δ. $\diamond$

In what follows, we tackle the problem of finding an accurate class $\Pi = \{\pi_\delta\}_{\delta\in(0,1]}$ of policies that minimizes the expected number of votes $\mathbb{E}_{f_*}^{\pi_\delta}[\tau]$ asymptotically as $\delta \downarrow 0$.

5. Lower and Upper Bounds on the Required Number of Votes

In this section, we derive upper and lower bounds for the sample complexity (i.e., the required number of votes) of any δ-accurate policy. Our lower bound result is rather general, as we provide an instance-specific and non-asymptotic lower bound on $\mathbb{E}_f^\pi[\tau]$ for any $f \in \mathcal{M}_p$ and any δ-accurate policy π. On the other hand, our upper bound is derived in an asymptotic regime under more restrictive conditions on the set of feasible preferences.

Let us start by introducing some notation. Given any display set $S \in \mathcal{S}$ and probability distribution $\lambda \in \Delta(\mathcal{S})$, the Kullback-Leibler (KL) divergence between two preferences f_1 and f_2 with respect to S and λ are given by

$$D_S(f_1||f_2) := \sum_{k\in S} f_1(k|S) \log \frac{f_1(k|S)}{f_2(k|S)} \qquad \text{and} \qquad D_\lambda(f_1||f_2) := \sum_{S\in\mathcal{S}} \lambda(S)\, D_S(f_1||f_2), \tag{1}$$

respectively. We define $\mathcal{M}_p(f) := \{f' \in \mathcal{M}_p : \sigma_f(1) = \sigma_{f'}(1)\}$ to be the set of preferences that have the same top-ranked version as f and its complement $\overline{\mathcal{M}}_p(f) := \mathcal{M}_p \setminus \mathcal{M}_p(f)$, which is the set of preferences that have a top-ranked version different from f. We also introduce *Chernoff's information measure* $I_*(f)$ to be the value of the following Max-Min problem (parameterized by preference f):

$$I_*(f) := \sup_{\lambda\in\Delta(\mathcal{S})} \inf_{\bar{f}\in\overline{\mathcal{M}}_p(f)} D_\lambda\left(f||\bar{f}\right). \tag{Max-Min}$$

Here, $I_*(f)$ quantifies the inherent difficulty of the learning problem when the underlying preference is f. As we will see below, $I_*(f)$ is a measure of separability among alternative hypotheses. In particular, the larger the value of $I_*(f)$, the easier the top-ranked identification problem is under f. Chernoff's information measure $I_*(f)$ is closely related to the expected number of votes needed by any δ-accurate policy. We use this relationship to calibrate the design of an optimal policy (see Section 7).

In passing, we note that the max-min nature of $I_*(f)$ allows for a game-theoretic interpretation. Under this interpretation, the decision-maker selects a randomized display strategy $\lambda \in \Delta(\mathcal{S})$ to maximize the KL divergence between a preference f and an alternative preference $\bar{f}$, which is being selected in an adversarial fashion from the set of preferences $\overline{\mathcal{M}}_p(f)$ that differ with f on the top-ranked product.

Given any $\delta_1, \delta_2 \in (0,1)$, let $kl(\delta_1, \delta_2) := \delta_1 \log \frac{\delta_1}{\delta_2} + (1-\delta_1) \log \frac{1-\delta_1}{1-\delta_2}$ denote the Kullback-Leibler divergence between two Bernoulli distributions with means δ_1 and δ_2. Theorem 1 below identifies a lower bound on the number of votes needed by any δ-accurate policy.

Theorem 1. (Lower Bound on $\mathbb{E}^{\pi}_{\sigma}[\tau]$) *Let $\delta \in (0,1)$. For any δ-accurate policy π and $f \in \mathcal{M}_p$,*

$$\mathbb{E}^{\pi}_{f}[\tau] \geq \frac{kl(\delta, 1-\delta)}{I_*(f)} \qquad \textit{and} \qquad \liminf_{\delta \downarrow 0} \frac{\mathbb{E}^{\pi}_{f}[\tau]}{\log\left(\frac{1}{\delta}\right)} \geq \frac{1}{I_*(f)}.$$

Our proof of Theorem 1 in Appendix B is based on Kaufmann et al. (2016) and Garivier and Kaufmann (2016). Specifically, we adapt the change of measure Lemmas 18 and 19 in Kaufmann et al. (2016) that they develop for the best arm identification setting to a more general hypothesis testing framework that we describe in Appendix B.1, which captures our setting for identifying the top-ranked alternative as a special case. In the process, we also provide a slight generalization of the 'transportation' Lemma 1 in Kaufmann et al. (2016). This extra level of generality in the proof of Theorem 1 reveals that the lower bound above can be applied to a broader class of problems such as identifying the top-k versions or the complete ranking σ_{f_*}, to name a few. We illustrate this point in Section B.5 in the appendix, where we compare our lower bound to the one proposed by Jamieson et al. (2015) (see also Heckel et al. 2019) in the context of dueling bandits. In Proposition 7 we show that our bound is tighter than the one proposed by Jamieson et al. (2015) (Theorem 3 in their paper).

Let us now turn to the derivation of the upper bound. As we mentioned above, to derive this upper bound we need to impose some additional conditions. Specifically, we need to limit the cardinality of the set of feasible preferences.

Theorem 2. (Upper Bound on $\mathbb{E}^{\pi}_{\sigma}[\tau]$) *Let $\mathcal{M}^{\mathrm{F}}_p$ be an arbitrary finite subset of $\mathcal{M}_p$ and $\delta \in (0,1)$. There exists a $\delta(\mathcal{M}^{\mathrm{F}}_p)$-accurate policy $\hat{\pi}$ such that $\mathbb{E}_f[\tau] < \infty$ for every $\delta \in (0,1)$ and*

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}^{\hat{\pi}}_{f}[\tau]}{\log\left(\frac{1}{\delta}\right)} \leq \frac{1}{I_*(f)}.$$

The proof of Theorem 2 is in Appendix C. Asymptotically speaking, Theorems 1 and 2 imply that $\mathbb{E}_f[\tau] \approx \frac{\log \frac{1}{\delta}}{I_*(f)}$ under any $\delta(\mathcal{M}^{\mathrm{F}}_p)$-accurate policy and for any $f \in \mathcal{M}^{\mathrm{F}}_p$. Since the set of preferences $\mathcal{M}^{\mathrm{F}}_p$ is arbitrary (except for the finiteness requirement), this relationship suggests that the dependence between consumers' preference f and the speed of learning of any δ-accurate policy can be quantified by Chernoff's information measure $I_*(f)$. In particular, we expect that the larger the value of $I_*(f)$ the faster one can learn the underlying preference f. Since our goal is to find a δ-accurate policy that minimizes the amount of time needed to learn the top-ranked version uniformly over the class of p-Separable choice models, the objective now turns to identifying the policy that performs well against the p-Separable choice model with the smallest value of $I_*(f)$.

In the next section, we characterize this hardest-to-learn consumer choice model (which we refer to it as the *Ordinal Attraction* choice model or OA) by solving a robust version of the Max-Min problem above. Then, in Section 7, we propose a δ-accurate policy that (asymptotically) achieves the lower bound in Theorem 1 and satisfies $\mathbb{E}_f[\tau] \approx \frac{\log \frac{1}{\delta}}{I_*(f)}$ under the OA model.

6. Worst-Case Analysis: Ordinal Attraction Model

Motivated by the lower bound in Theorem 1, in this section we find a subset of preferences in $\mathcal{M}_p$ that all have the smallest value of $I_*(f)$. To this end, let

$$I_*^{\mathrm{OA}} := \inf_{f \in \mathcal{M}_p} I_*(f). \tag{2}$$

The reason for the superscript "OA" defining I_*^{OA} will become apparent in Theorem 3 below. If f is any preference such that $I_*(f) = I_*^{\mathrm{OA}}$ then any permutation of f also minimizes $I_*(f)$ since the labeling of the K versions is completely arbitrary.[7]

A subset of $\arg\min_{f \in \mathcal{M}_p} I_*(f)$ is described in Theorem 3 below. These minimizing preferences make use of the following definition:

$$\sigma(X|S) := \sum_{k \in S} \mathbb{I}\{\sigma(k) \leq \sigma(X)\} \quad \text{for any } X \in S.$$

That is, $\sigma(\cdot|S) : S \to [|S|]$ is the restriction of σ to S so that $\sigma(k_1|S) < \sigma(k_2|S)$ if and only if $\sigma(k_1) < \sigma(k_2)$ for every $k_1, k_2 \in S$.

Theorem 3. *Let $p \in [0,1)$. For every $\sigma \in \Sigma$ define the preference*

$$f_\sigma^{\mathrm{OA}}(X|S) := \frac{1-p}{1-p^{|S|}} p^{\sigma(X|S)-1} \qquad S \in \mathcal{S} \text{ and } X \in S. \tag{3}$$

and let

$$\mathcal{M}_p^{\mathrm{OA}} := \{f_\sigma^{\mathrm{OA}} : \sigma \in \Sigma\}. \tag{4}$$

Then $\mathcal{M}_p^{\mathrm{OA}} \subseteq \arg\min_{f \in \mathcal{M}_p} I_*(f)$.

The proof of Theorem 3 is in Appendix D.

Let $\overline{\mathcal{M}}_p^{\mathrm{OA}}(f) := \{\bar{f} \in \mathcal{M}_p^{\mathrm{OA}} : \sigma_{\bar{f}}(1) \neq \sigma_f(1)\}$ be the class of OA preferences that disagree with f on the top-ranked version. The following result follows from the proof of Theorem 3.

Corollary 1. *For any $f \in \mathcal{M}_p^{\mathrm{OA}}$ we have that*

$$I_*^{\mathrm{OA}} = I_*(f) = \max_{\lambda \in \Delta(\mathcal{S})} \min_{\bar{f} \in \overline{\mathcal{M}}_p^{\mathrm{OA}}(f)} D_\lambda\left(f || \bar{f}\right).$$

[7] Indeed, given any $f \in \mathcal{M}_p$ and any permutation $\sigma \in \Sigma$, let $f_\sigma \in \mathcal{M}_p$ be defined by $f_\sigma(X|S) = f(\sigma(X)|\sigma(S))$ for all $X \in [K]$ and $S \in \mathcal{S}$. Then, it is not hard to see that $I_*(f) = I_*(f_\sigma)$.

In other words, when computing the value of I_*^{OA}, we can replace the set of alternative preferences $\overline{\mathcal{M}}_p(f)$ in the Max-Min problem by the considerably smaller set of alternatives OA preferences $\overline{\mathcal{M}}_p^{\text{OA}}(f)$. As we will see, this reduces significantly the complexity of characterizing the optimal randomization strategy $\lambda \in \Delta(\mathcal{S})$. We will also exploit Corollary 1 to formulate a robust version of the company's problem.

Discussion of the set of OA preferences $\mathcal{M}_p^{\text{OA}}$. To get some intuition about the structure of a preference $f_\sigma^{\text{OA}} \in \mathcal{M}_p^{\text{OA}}$, note that Theorem 3 implies that the likelihood that a consumer selects version X out of the display set S is proportional to the relative ranking of product X within S. In other words, the "attractiveness" of product X has an ordinal dependence on the set S. It is because of this property that we refer to f_σ^{OA} as an *Ordinal Attraction* (OA) choice model. Mathematically, it is not hard to see that f_σ^{OA} satisfies the p-Separability requirement (A-4) in Definition 1 with equalities.

Although we derived the OA preferences for the purpose of characterizing the smallest value of Chernoff's information measure I_* within the class of p-Separable choice model, the OA model has a number of additional properties. For instance, the OA model is an extension of the popular class of noisy pairwise comparison models that have been widely used in the literature (e.g. Braverman and Mossel 2008, Braverman and Mossel 2009, Caragiannis et al. 2013, Wauthier et al. 2013, to name a few). As its name suggests, in a pairwise noisy comparison model, only pairs of products are displayed and the better version is chosen with a fixed probability independent of the actual pair being displayed. It is easy to see that when only pairs are displayed, the Ordinal Attraction model reduces to the pairwise comparison model.

From a practical standpoint, another appealing feature of the OA model is that it provides a parsimonious framework to study consumer choice behavior using limited information about product version characteristics. Indeed, by its ordinal nature, the only relevant attribute of a version for the purpose of affecting consumers' voting choices is its relative ranking within the display set.

Finally, as we will see below, the OA model is also tractable and allows us to derive a complete closed-form characterization of our proposed Myopic Tracking algorithm. We leverage this solution to derive valuable insights into the structure of an optimal policy and how to effectively balance the coverage-accuracy trade-off discussed in the Introduction.

Computing the Information Measure I_*^{OA}. Using the result in Theorem 3 and Corollary 1, let us now turn to the question of determining the lower bound for Chernoff's information measure I_* that solves the Max-Min problem above. That is, for any $f_\sigma^{\text{OA}} \in \mathcal{M}_p^{\text{OA}}$, we are interested in computing

$$I_*^{\text{OA}} = \max_{\lambda \in \Delta(\mathcal{S})} \min_{\bar{f} \in \overline{\mathcal{M}}_p(f_\sigma^{\text{OA}})} D_\lambda\left(f_\sigma^{\text{OA}} || \bar{f}\right).$$

We note that the specific ranking σ that we use in this definition is immaterial since the preference f_σ^{OA} is only defined up to permutations of σ (see footnote 7). So, to simplify the notation, we will assume that σ is equal to the consensus ranking $\sigma_*(k) = k$ for all $k \in [K]$ and denote $f_*^{\mathrm{OA}} = f_{\sigma_*}^{\mathrm{OA}}$. Our next result characterizes optimal solutions for the randomized display strategy problem

$$\max_{\lambda \in \Delta(\mathcal{S})} \min_{\bar f \in \overline{\mathcal{M}}_p(f_*^{\mathrm{OA}})} D_\lambda\left(f_*^{\mathrm{OA}} || \bar f\right). \tag{5}$$

Theorem 4. *Define the sequences* $\{\mathfrak{a}_n\}$, $\{\mathfrak{b}_n\}$ *and* $\{\lambda_n^*\}$ *of positive real numbers as follows:*

$$\mathfrak{a}_n := \log\left(\tfrac{1}{p}\right)\left[1 - np^{n-1} + (n-1)p^n\right], \quad \mathfrak{b}_n = 1 - p^n, \quad \textit{and} \quad \lambda_n^* = \begin{cases} \mathfrak{b}_n\left(\frac{1}{\mathfrak{a}_n} - \frac{1}{\mathfrak{a}_{n+1}}\right), & \textit{if } n = 2, \ldots, K-1; \\ \frac{\mathfrak{b}_n}{\mathfrak{a}_n}, & \textit{if } n = K. \end{cases} \tag{6}$$

Then, the unique optimal solution (of the outer maximization) of (5) *is given by*

$$\lambda_*^{\mathrm{OA}}(S) = \begin{cases} \frac{\lambda_n^*}{\lambda_2^* + \cdots + \lambda_K^*} & \textit{if } S = [n] \textit{ for some } n \in \{2, \ldots, K\} \\ 0 & \textit{otherwise.} \end{cases} \tag{7}$$

The result in Theorem 4 is significant for a number of reasons. First, it shows that the randomized display rule in (7) is static, independent of the voting history, and can be computed offline. Second, it reveals a nested structure of the display sets that have a positive probability of being offered. Indeed, note that the support of λ_*^{OA} is the collection of display sets: $\big\{[2], [3], \ldots, [K]\big\}$. Finally, this collection is rather sparse (there are only $K-1$ members out of the $2^K - 1$ possible display sets) which is a fact that significantly simplifies its computation and implementation.

Equipped with Theorem 4, we can now compute the smallest value I_*^{OA} of $I_*(f)$ within the class of p-Separable preferences.

Proposition 2. *The value of* I_*^{OA} *is*

$$I_*^{\mathrm{OA}} = (1-p)\log\left(\tfrac{1}{p}\right)\left(1 + \sum_{n=2}^{K} \frac{p^{n-1}}{1 + 2p + \cdots + (n-1)p^{n-2}}\right)^{-1}.$$

It follows that $(1-p)\log\left(\frac{1}{p}\right) K\,(K + 2p(K-1))^{-1} \le I_*^{\mathrm{OA}} \le (1-p)\log\left(\frac{1}{p}\right)(1+p)^{-1}$.

From Proposition 2, one can see that I_*^{OA} decreases in both K and p. Also, numerical computations show that I_*^{OA} is not particularly sensitive to the value of K (the total number of versions). On the other hand, the impact of the noise parameter p is roughly given by $I_*^{\mathrm{OA}} \approx (1-p)^2$ as $p \uparrow 1$ and $I_*^{\mathrm{OA}} \approx \log\frac{1}{p}$ as $p \downarrow 0$.

7. Robust Learning and Myopic Tracking Policy

In this section, we propose a particular δ-accurate display policy, which we refer to it as *Myopic Tracking Policy* (MTP), and show that it is worst-case asymptotically optimal in a sense that we make precise in Theorem 5.

Inspired by the discussion in the previous section, the implementation of the Myopic Tracking Policy leverages the structure of the OA model. More specifically, it has three main steps: (1) an MLE estimation of the true ranking (*i.e.*, most likely hypothesis) restricted to the OA model, (2) a stopping criteria to decide whether to stop the voting process or to continue it, given the available information, and (3) a random selection of the display set to show to the next consumer in case the voting process is continued. The specific details of the Myopic Tracking policy are provided in Algorithm 1 below, whose statement makes use of the following definitions. First, given any history $H_t = (S_1, X_1, \dots, S_t, X_t)$ and any pair of preferences $f, \bar{f} \in \mathcal{M}_p$, we define the log-likelihood ratio process

$$L_t^{f,\bar{f}} := \sum_{\ell=1}^{t} \log\left(\frac{f(X_\ell|S_\ell)}{\bar{f}(X_\ell|S_\ell)}\right). \tag{8}$$

Second, for any preference $f \in \mathcal{M}_p$, we define a nested sequence of display sets $\{S_f(k) \colon k = 2, \dots, K\}$ such that $S_f(k) := \{\sigma_f^{-1}(\ell) \colon \ell = 1, \dots, k\}$ is the set of top-k versions under f.

Algorithm 1 is parameterized by a single exogenous parameter β (possibly depending on the error tolerance δ) that is used in the stopping criterion in Step 2. The choice of the parameter β is critical to ensure that the Myopic Tracking policy is both fast and δ-accurate. We discuss these two issues next.

7.1. Worst-case Asymptotic Optimality of the Myopic Tracking Policy.

We now show that the Myopic Tracking Policy is worst-case asymptotically optimal for an appropriate choice of the parameter $\beta = \beta(\delta)$ as a function of δ. To formalize this result, recall that a family of policies $\Pi = \{\pi_\delta\}_{\delta \in (0,1]}$ is *accurate* if each member π_δ is δ-accurate.

THEOREM 5. (Worst-case asymptotic optimality of MTP) *There exists a threshold $\beta = \beta(\delta)$ such that*

$$\text{MTP} \in \underset{\Pi \textit{ is accurate}}{\arg\min} \; \sup_{f \in \mathcal{M}_p} \; \limsup_{\delta \downarrow 0} \; \frac{\mathbb{E}_f^{\pi_\delta}[\tau]}{\log(1/\delta)}. \tag{9}$$

The proof of Theorem 5 consists of two parts. First, in Theorem 6, we show how to select β so that MTP uses an expected number of votes that asymptotically matches the lower bound in Theorem 1. Second, in Theorem 7, we show how to select β to ensure that MTP belongs to the family of δ-accurate policies.

Both the stopping time τ and threshold β depend on the error tolerance δ. More specifically, because of the hitting time property (see Figure 1), τ is an increasing function of β, and let us

Algorithm 1 Myopic Tracking Policy (MTP)

```
INPUT: A scalar $\beta = \beta(\delta) > 0$.
STEP 0: (Initialization). Set $t = 1$, select an arbitrary display set $S_1 \in \mathcal{S}$ to show to the first
        consumer and record the vote $X_1$.
STEP 1: At time epoch $t$, given the history of votes $(S_1, X_1, \ldots, S_t, X_t)$, compute a most
        likely consensus preference by solving the MLE problem
            $f_t^{\mathrm{OA}} \in \arg\max_{f \in \mathcal{M}_p^{\mathrm{OA}}} \sum_{\ell=1}^{t} \log f(X_\ell | S_\ell).$    (MLE)
        We break ties uniformly at random if the arg max in (MLE) is not a singleton.
STEP 2: Update the value of the generalized log-likelihood ratio process
            $\mathcal{L}_t = \min_{\bar{f} \in \overline{\mathcal{M}}_p^{\mathrm{OA}}(f_t^{\mathrm{OA}})} L_t^{f_t^{\mathrm{OA}}, \bar{f}}.$    (L)
        If $\mathcal{L}_t \geq \beta$, then stop set $\tau = t$ and select the top-ranked version according to $f_t^{\mathrm{OA}}$;
        that is, $\sigma_{f_t^{\mathrm{OA}}}^{-1}(1)$. Otherwise, go to Step 3.
STEP 3: For $k = 2, \ldots, K$, let $\hat{S}_t(k) = S_{f_t^{\mathrm{OA}}}(k)$ and $\hat{\lambda}_t(k) = \lambda_*^{\mathrm{OA}}([k])$ (see Theorem 4).
        Using the vector of probabilities $\hat{\lambda}_t(k)$, randomly select from the set
        $\{\hat{S}_t(2), \hat{S}_t(3), \ldots, \hat{S}_t(K)\}$ the next display set $S_{t+1}$ to be displayed to the next con-
        sumer and record her choice $X_{t+1}$. Go to Step 1 and iterate. □
```

call this dependence $\tau(\beta)$. Meanwhile, we also expect that the lower the tolerance δ is, the higher β needs to be. Let us denote this dependence as $\beta(\delta)$. Intuitively, β is a decreasing function of δ. Combining $\tau(\beta)$ and $\beta(\delta)$, we expect $\tau = \tau(\beta(\delta))$ to decrease in δ.

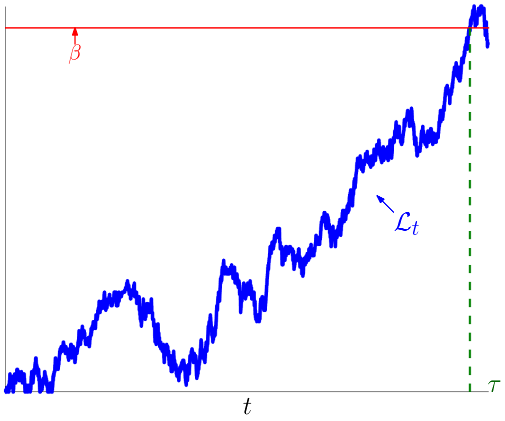


Figure 1 Illustration of the Myopic Tracking Policy. The stopping time τ is a hitting time.

Theorem 6 below gives a sufficient condition on how "small" $\beta = \beta(\delta)$ needs to be for the Myopic Tracking Policy to be fast in the sense of Theorem 1.

THEOREM 6. (Sample Complexity of MTP) *For any constant C_0 (independent of δ), threshold $\beta = \beta(\delta)$ such that $\beta \le C_0 + \log \frac{1}{\delta}$ and preference $f \in \mathcal{M}_p$, we have that $\mathbb{E}_f[\tau] < \infty$ for every $\delta > 0$ under the Myopic Tracking Policy. Moreover,*

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_f[\tau]}{\log\left(\frac{1}{\delta}\right)} \le \frac{1}{I_*^{\mathrm{OA}}}. \tag{10}$$

The inequality above can be replaced by an equality when $f \in \mathcal{M}_p^{\mathrm{OA}}$.

Theorem 6 implies that the MTP uses a number of votes that asymptotically matches the lower bound of any δ-accurate policy. The key idea behind the proof of Theorem 6 is that the MTP matches the randomization strategy λ_*^{OA} in Theorem 4. As a result, the stochastic process $\mathcal{L}_t$ achieves the fastest rate of growth (*i.e.*, maximum drift). We will expand more on this reasoning in Section 7.3.

Our next result gives a sufficient condition on how "large" $\beta = \beta(\delta)$ needs to be for the Myopic Tracking Policy to be δ-accurate.

THEOREM 7. (Accuracy of MTP) *There exists a constant C_1, that depends only on K, such that as long as $\beta \ge C_1 + \log\left(\frac{1}{\delta}\right)$, the Myopic Tracking policy is δ-accurate for every $\delta \in (0,1)$.*

The hitting threshold β controls the error probability of the MTP algorithm. Intuitively, the higher the threshold the less likely it is that we will end up selecting the wrong version in the end. The proof of Theorem 7 is based on a change-of-measure argument and does not rely on the specific structure of the MTP (except the requirement that $\tau < \infty$ $\mathbb{P}_f$−a.s. for any $f \in \mathcal{M}_p^{\mathrm{OA}}$). The construction of threshold β is based on an estimate of the error probability $\mathbb{P}_f(d_\tau \neq \sigma_f^{-1}(1))$ for every $f \in \mathcal{M}_p$. The main idea behind the estimation is two-fold: first, by tracking the generalized log-likelihood process restricted to the OA model, the MTP achieves δ-accuracy within the OA model (based on a change-of-measure argument); second, since OA model is the "hardest to learn", achieving δ-accuracy within the OA model also implies achieving δ-accuracy within $\mathcal{M}_p$ (based on a dominance argument).[8]

Finally, we can combine the results in Theorems 1, 6, and 7 to establish the worst-case asymptotic optimality of the MTP.

[8] Chernoff (1959) proposed a similar type of hitting threshold for an alternative performance criterion. His analysis, however, assumed that the number of states (i.e., preferences in our setting) is finite and does not extend to our case in which the cardinality of $\mathcal{M}_p$ is infinite. More recently, in the context of best-arm identification, Garivier and Kaufmann (2016) have also proposed a sampling policy that relies on a similar hitting threshold like the one in Theorem 7. However, because of the structure of their problem, to ensure δ-accuracy their threshold must grow at a rate of $\log(t)$ over time while in our case the threshold is a constant independent of t.

Proof of Theorem 5. Select an arbitrary δ-accurate policy π. Theorem 1 and (2) imply

$$\sup_{f\in\mathcal{M}_p} \limsup_{\delta\downarrow 0} \frac{\mathbb{E}_f^{\pi}[\tau]}{\log(1/\delta)} \geq \sup_{f\in\mathcal{M}_p} \frac{1}{I_*(f)} = \frac{1}{I_*^{\mathrm{OA}}}.$$

By Theorem 7, MTP is δ-accurate if we pick $\beta = C_1 + \log\left(\frac{1}{\delta}\right)$. Invoking Theorem 6 with that β yields

$$\sup_{f\in\mathcal{M}_p} \limsup_{\delta\downarrow 0} \frac{\mathbb{E}_f^{MTP}[\tau]}{\log(1/\delta)} \leq \sup_{f\in\mathcal{M}_p} \frac{1}{I_*^{\mathrm{OA}}} = \frac{1}{I_*^{\mathrm{OA}}}.$$

As a result, (9) holds. ■

Remark 3. The conditions in Theorem 6 and 7 may be further weakened. For example, f_t^{OA} defined in (MLE) does not have to be the maximum likelihood estimator, but any statistic such that:

$$\mathbb{E}_{f_*^{\mathrm{OA}}}[\hat{\tau}^2] < \infty, \text{ where } \hat{\tau} := \max\{t : f_t^{\mathrm{OA}} \neq f_*^{\mathrm{OA}}\} \tag{11}$$

Here $\hat{\tau}$ is a $\mathbb{Z}_+ \cup \{+\infty\}$-valued, $\mathcal{F}_\infty$-measurable[9] random variable that denotes the last time period in which the estimated preference f_t^{OA} is not equal to f_*^{OA}. ◇

7.2. On the Complexity of the Myopic Tracking Policy

In this subsection, we discuss the computational complexity of the Myopic Tracking Policy. Each iteration of Algorithm 1 involves three steps. Because of the result in Theorem 4, the third step is rather simple as it involves randomly selecting a display set out of $K-1$ alternatives. Since the randomization probabilities are fixed, independent of the history of the learning process, this step can be executed offline before the feedback process begins. Steps 1 and 2, on the other hand, involve solving a possibly large combinatorial optimization problem at each iteration. Fortunately, the structure of the underlying OA preference model allows us to solve these combinatorial problems without much computational burden. The rest of this section is devoted to supporting this claim.

To this end, let us represent the voting history $H_t = (S_1, X_1, \ldots, S_t, X_t)$ using a complete directed graph $\mathcal{G}_K$ with K nodes, each representing a version. For each arc $(i,j) \in [K]\times[K]$, we define its weight $w_{i,j}^t$ by

$$w_{ij}^t := \sum_{\ell=1}^{t} \mathbb{I}\big\{\{i,j\} \subseteq S_\ell \text{ and } X_\ell = i\big\}. \tag{12}$$

That is to say, w_{ij}^t is the total number of instances, up to time t, where both versions i and j are jointly displayed and version i is voted for. Intuitively, if $w_{ij}^t - w_{ji}^t$ is large then there is a strong indication that version i has a higher ranking than version j.

[9] $\mathcal{F}_\infty$ is defined as the smallest sigma-algebra containing $\cup_{t=1}^{\infty}\mathcal{F}_t$.

We use the graph $\mathcal{G}_K$ to quantify the discrepancy between any given preference f and the voting history using the total weight of the *feedback arc set* $\{(i,j):\sigma_f(j)<\sigma_f(i)\}$. We introduce the discrepancy cost

$$c(f,\vec{w}^t):=\sum_{(i,j):i\neq j}\mathbb{I}\{\sigma_f(j)<\sigma_f(i)\}\,w^t_{ij}. \tag{13}$$

The argument inside the summation is the total number of instances where a pair of versions (i,j) are jointly displayed and the less preferred version i under f (that is, $\sigma_f(i)>\sigma_f(j)$) is chosen.

In the result below, we demonstrate that the log-likelihood of any preference f given voting history H_t is proportional to its discrepancy cost $c(f,\vec{w}^t)$. As a result, (MLE) (resp. (L)) corresponds to an unconstrained (resp. constrained) discrepancy cost minimization problem.

Proposition 3. *Given the voting history $H_t=(S_1,X_1,\ldots,S_t,X_t)$, the following facts hold:*

1. *There exists a constant ϕ such that for any $f\in\mathcal{M}_p^{\mathrm{OA}}$, $\sum_{\ell=1}^t\log f(X_\ell;S_\ell)=\log(p)\cdot c(f,\vec{w}^t)+\phi$. Hence* (MLE) *is equivalent to finding an*

$$f_t^{\mathrm{OA}}\in\arg\min_{f\in\mathcal{M}_p^{\mathrm{OA}}}c(f,\vec{w}^t).$$

2. *Given any $f,\bar f\in\mathcal{M}_p^{\mathrm{OA}}$, $L_t^{f,\bar f}=\log\left(\frac{1}{p}\right)\cdot d(f,\bar f)$, where $d(\cdot,\cdot)$ is defined as*

$$d(f,\bar f):=c(\bar f,\vec{w}^t)-c(f,\vec{w}^t)=\sum_{(i,j):i\neq j}(w^t_{i,j}-w^t_{j,i})\,\mathbb{I}\{\sigma_{\bar f}(j)<\sigma_{\bar f}(i)\ \textit{but}\ \sigma_f(j)>\sigma_f(i)\}. \tag{14}$$

Hence (L) *is equivalent to solving*

$$\mathcal{L}_t=\min_{\bar f\in\overline{\mathcal{M}}_p^{\mathrm{OA}}(f_t^{\mathrm{OA}})}d(f_t^{\mathrm{OA}},\bar f).$$

Proposition 3 reveals that the (MLE) problem in Algorithm 1 is computationally equivalent to the *weighted feedback arc set problem on tournaments* on graph $\mathcal{G}_K$ with weights $\vec{w}^t$ (Ailon et al. 2005). This type of problems has been extensively studied in the computer science literature (Davenport and Kalagnanam 2004, Ailon et al. 2005, Alon 2006, Conitzer et al. 2006, Charbit et al. 2007, Kenyon-Mathieu and Schudy 2007, Schalekamp and Zuylen 2009, Fomin et al. 2010, etc.) and operations research literature (Grötschel et al. 1984, Mitchell and Borchers 1996, Charon and Hudry 2010). For example, acceleration algorithms are available for the following integer programming formulation of this problem (Grötschel et al. 1984): $\sigma_{\hat f}(i)=\sum_{j\in[K]\setminus\{i\}}\hat x_{ji}+1$, where:

$$\begin{aligned}\hat x\in\arg\min_{\vec x}&\sum_{(i,j):i\neq j}x_{ji}w^t_{ij}\\ s.t.\quad & x_{ij}+x_{jk}+x_{ki}\geq 1, && \forall\text{ distinct } i,j,k\in[K]\\ & x_{ij}+x_{ji}=1, && \forall\text{ distinct } i,j\in[K]\\ & x_{ij}\in\{0,1\}. && \forall\text{ distinct } i,j\in[K]\end{aligned} \tag{15}$$

It has also been reported that the integer programming formulation is effective both on randomly generated data sets (Conitzer et al. 2006) and on real data sets (Ali and Meila 2012). Moreover, Kenyon-Mathieu and Schudy (2007) propose a polynomial-time approximation scheme (PTAS) to approximate the problem. That is, when the number of versions is large and computational tractability is a concern, for any fixed ϵ, a polynomial-time algorithm with $1-\epsilon$ optimality is attainable. Note also that (L) can be reduced to (MLE) in linear time: for any $k \in [K] \setminus \{\sigma_{\hat{f}}^{-1}(1)\}$, we can solve for the most likely ranking conditional on k being top ranked. To do that, we just need to solve a weighted feedback arc set problem with sub-graph $\mathcal{G}_K \setminus \{k\}$, and insert k back to the ranking. We can solve (L) by comparing the optimal costs for all the sub-problems.

7.3. Discussion of the Myopic Tracking Policy

We conclude this section with a brief discussion of some key features of our proposed Myopic Tracking Policy. Let us start by providing some intuition behind the (asymptotic) optimality of MTP both in terms of the stopping rule and the choice of a display policy.

On the optimality of the stopping rule. To get some intuition behind the stopping rule used in the Myopic Tracking policy, we note that if the display policy $\{S_t\}$ is fixed, our problem reduces to a classical sequential (composite and multi-hypothesis) testing problem.[10] In this case, the idea of tracking the generalized log-likelihood ratio process $\mathcal{L}_t$, and stopping when $\mathcal{L}_t$ hits a pre-specified threshold, is known to be asymptotically optimal under various settings (e.g., Wald 1973, Chernoff 1972, Draglia et al. 1999, Li et al. 2014). Steps 1 and 2 in the Myopic Tracking policy extend this idea to our more general setting.

On the optimality of the display set policy. Given the (asymptotic) optimality of the sequential likelihood ratio test discussed above, the goal of an optimal display rule is to speed up the learning process by minimizing the time it takes the likelihood ratio process $\mathcal{L}_t$ to hit the threshold β. In other words, to solve the problem

$$\inf_{\{S_t\}} \mathbb{E}^{\pi}_{f_*} \inf_t \left\{ t : \mathcal{L}_t \geq \beta \right\}. \tag{16}$$

To build some intuition on the growth of $\mathcal{L}_t$, let us pick f_* from the OA model. Otherwise, $\mathcal{L}_t$ will grow more quickly, reflecting the fact that the company's learning problem is easier. Without loss of generality, let us pick $f_* = f^{\text{OA}}_{\sigma_*}$ within the OA model. The Myopic Tracking Policy is able to recover the ranking represented by f_*, or in other words, f_t^{OA} is absorbed into the preference f_*^{OA} quickly (in the sense of Equation (11)). Hence $\mathcal{L}_t = \min_{\bar{f} \in \overline{\mathcal{M}}_p^{\text{OA}}(f_t^{\text{OA}})} L_t^{f_t^{\text{OA}}, \bar{f}}$ is well-approximated

[10] In our problem, each hypothesis corresponds to a version being the top-ranked one, and each hypothesis contains the family of rankings that rank the same version at the top.

by the process $\min_{\bar{f}\in\overline{\mathcal{M}}_p^{\text{OA}}(\sigma_*)} L_t^{f_*^{\text{OA}},\bar{f}}$ as t grows. Further, we may understand the latter process to consist of two components: a deterministic part (which is a deterministic process that grows linearly) and a noise part (which is a random process that diverges sub-linearly). The deterministic part captures the growth rate of $\mathcal{L}_t$ and could be written in the following manner:

$$\tilde{\mathcal{L}}_t := \min_{\bar{f}\in\bar{f}^{\text{OA}}} \sum_{\ell=1}^{t} \sum_{k\in S_\ell} f_*(k|S_\ell) \log\left(\frac{f_*(k|S_\ell)}{\bar{f}(k|S_\ell)}\right) = \min_{\bar{f}\in\overline{\mathcal{M}}_p^{\text{OA}}(f_*)} \sum_{\ell=1}^{t} D_{S_\ell}\left(f_*||\bar{f}\right)$$
$$= t\cdot \underbrace{\min_{\bar{f}\in\overline{\mathcal{M}}_p^{\text{OA}}(f_*)} D_{\bar{\lambda}}\left(f_*||\bar{f}\right)}_{\text{growth rate}}, \quad \text{where } \bar{\lambda}(S) = \underbrace{\frac{\sum_{l=1}^{t}\mathbb{I}\{S_l = S\}}{t}}_{\text{display frequency of set } S}.$$

As a result, by replacing $\mathcal{L}_t$ with $\tilde{\mathcal{L}}_t$, we can replace the optimal hitting problem in (16) by the problem of maximizing the average growth rate of $\tilde{\mathcal{L}}_t$. Furthermore, to maximize this average growth rate, it suffices to balance the long-run display frequency of each set to achieve the fastest growth rate of $\tilde{\mathcal{L}}_t$ (see Figure 2). This can be done by selecting

$$\bar{\lambda} \in \arg\max_{\lambda\in\Delta(\mathcal{S})} \min_{\bar{f}\in\overline{\mathcal{M}}_p^{\text{OA}}(f_*)} D_\lambda\left(f_*||\bar{f}\right).$$

That is, selecting $\bar{\lambda}$ that solves (Max-Min) with preference f_*, so that

$$\tau \approx \beta / \max_{\lambda\in\Delta(\mathcal{S})} \min_{\bar{f}\in\overline{\mathcal{M}}_p^{\text{OA}}(f_*)} D_\lambda\left(f_*||\bar{f}\right) = \beta / I_*(f_*).$$

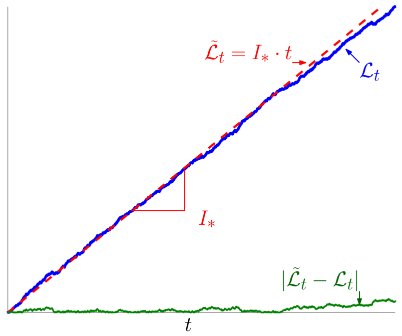


Figure 2 Intuition behind the Myopic Tracking Policy. Over a long time, $\mathcal{L}_t$ is well-approximated by $\tilde{\mathcal{L}}_t$, a linear function. The display policy is chosen to maximize the slope of $\tilde{\mathcal{L}}_t$.

Accuracy versus Coverage. Recall that in the Introduction we have argued that an effective display policy should balance the underlying accuracy/coverage trade-off embedded in the problem. On the one hand, the company can show every consumer the entire set $[K]$ to use every vote to learn about every product. On the other hand, the quality of the information collected on each vote is typically higher when the cardinality of the display set is small. As we show next, the Myopic Tracking Policy resolves this tension in a rather parsimonious fashion.

Indeed, one of the key features of the MTP is the simplicity of its display policy (i.e, Step 3 in Algorithm 1). As mentioned above, the policy randomizes over a nested collection of $K-1$ display sets. Specifically, if f_t^{OA} is the MLE estimate of f_* at period t, then the MTP policy randomizes over the sets $\{S_{f_t^{\mathrm{OA}}}(k)\colon k=2,\dots,K\}$. (Recall that $S_{f_t^{\mathrm{OA}}}(k)$ is the set that includes the top-k versions under f_t^{OA}.) Furthermore, from Theorem 4, the probability of displaying set $S_{f_t^{\mathrm{OA}}}(k)$ is equal to

$$\lambda_*^{\mathrm{OA}}(S_{f_t^{\mathrm{OA}}}(k)) = \lambda_k^*/(\lambda_2^* + \cdots + \lambda_K^*) \qquad \text{for } k=2,\dots,K, \tag{17}$$

where the values of the $\{\lambda_k^*\}$ are given in the theorem. One can show that these randomization probabilities satisfy the following property.

Corollary 2. (U-shaped λ_*^{OA}) *For any $K \geq 4$, we have $\lambda_2^* > \lambda_3^* > \cdots > \lambda_{K-1}^*$ and $\lambda_{K-1}^* < \lambda_K^*$.*

In other words, the vector of randomization probabilities is “U” shaped as a function of the cardinality of the display sets; it decreases with the cardinality and then increases for the full display set $[K]$. The proof of Corollary 2 is in Appendix F. Figure 3 illustrates the values of $\{\lambda_k^*\}$ for different values of K and p. Roughly speaking, the proposed display strategy in the MTP policy exhibits the following pattern:

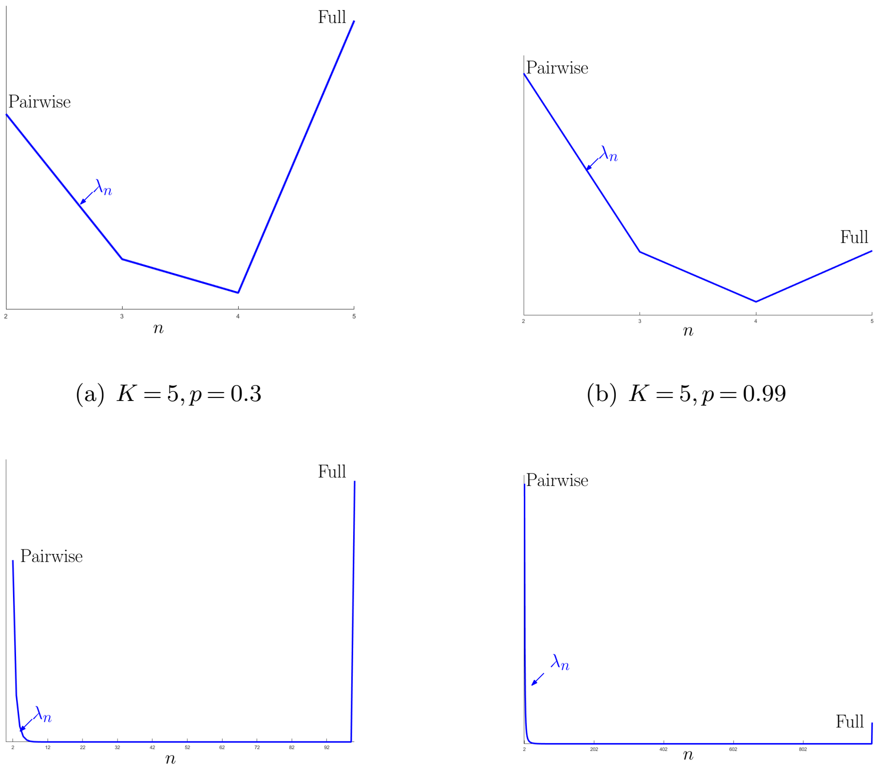
- The noisier the environment is (*i.e.,* p is larger), the more probability is allocated to smaller display sets. The less noisy the environment, the more probability is allocated to the full set (*i.e.,* full-display).
- As the number of versions grows (*i.e.,* K is larger), MTP tends to allocate larger probabilities to either very small display sets (and mostly pairwise) or full-display.
- For sufficiently large values of K, MTP appears to randomize only between pairwise and full-display.

To further underscore these patterns, Table 1 reports the values of

$$\underline{\Lambda}_i := \sum_{k=2}^{i} \lambda_k^* \qquad \text{and} \qquad \overline{\Lambda}_i = \sum_{k=K-i+1}^{K} \lambda_k^*,$$

which are the probabilities of selecting a display set with cardinality less than or equal to i or greater than or equal to $N-i+1$, respectively

As we can see from the table, for small values of p, $\underline{\Lambda}_2 + \overline{\Lambda}_2$ is almost one and so the MTP essentially uses two display sets: (i) a pair with the top two versions and (ii) the full set with all K versions, with more than 80% allocated to the full display set. The table also shows that as p grows the display policy of MTP relies on additional display sets of small cardinality. For instance, even



(a) $K = 5, p = 0.3$

(b) $K = 5, p = 0.99$

(c) $K = 100, p = 0.3$

(d) $K = 1000, p = 0.9$

Figure 3 Display probabilities λ_*^{OA} of Myopic Tracking Policy for different values of K and p. In each panel, n is the cardinality of the nested display set $\hat{S}(n)$.

(K,p)	$\underline{\Lambda}_2$	$\overline{\Lambda}_2$	$\underline{\Lambda}_5$	$\overline{\Lambda}_5$	$\underline{\Lambda}_{10}$	$\overline{\Lambda}_{10}$
(20,0.1)	0.165	0.811	0.189	0.811	0.189	0.811
(20,0.5)	0.440	0.293	0.671	0.293	0.705	0.294
(20,0.9)	0.468	0.053	0.790	0.065	0.897	0.094
(100,0.1)	0.165	0.811	0.189	0.811	0.189	0.811
(100,0.5)	0.440	0.293	0.671	0.293	0.705	0.293
(100,0.9)	0.465	0.038	0.781	0.038	0.891	0.038
(200,0.1)	0.165	0.811	0.189	0.811	0.189	0.811
(200,0.5)	0.440	0.293	0.671	0.293	0.705	0.293
(200,0.9)	0.465	0.038	0.781	0.038	0.891	0.038

Table 1 Λ_i: Probability that MTP selects a display set with cardinality less than or equal to i or greater than or equal to $N - i + 1$ for different values of K and p.

when $p = 0.9$ and consumer preferences are rather noisy, $\underline{\Lambda}_{10} \approx 90\%$ independent of K. In other words, if the company has $K = 200$ versions and use the MTP policy, about 90% of the display sets would include less than ten versions.

Pairwise versus Multiwise Comparisons. Proposition 3 reveals an interesting feature about how information accumulates under an OA choice model. It shows that it is *ex-post* equivalent to replace a vote from a multi-wise comparison with a collection of pairwise comparisons. For example, if we replace a vote on Version 1 from the display set $S = \{1,2,3\}$ with two consecutive votes on Version 1 from two pairs $\{1,2\}$ and $\{1,3\}$ respectively, we do not change the discrepancy cost or likelihood of any preference f. This property explains why we can use a simple graph representation to summarize the voting data. The idea of breaking a multi-wise comparison into several pairwise comparisons relates to Jiang et al. (2011).

Connection to Voting Theory. The discrepancy metric in (13) used to solve (MLE) and (L) resembles the one used in the Kemeny-Young method in voting theory (Young 1988, Levin and Nalebuff 1995). As a result, our choice model inherits a number of desirable properties of the Kemeny-Young model. For example, if there exists a version that gets more votes than any other version when they are jointly displayed (known as the *Condorcet winner*), that version is the optimal choice.

8. Numerical Experiments

In this section, we numerically investigate the performance of the Myopic Tracking Policy. First, in Section 8.1, we study the running time of Algorithm 1. In Section 8.2, we compare the sample complexity of MTP and three benchmark policies that use alternative display strategies.

8.1. Running Time of the Myopic Tracking policy

Most of the computational time required to implement the Myopic Tracking policy in Algorithm 1 is allocated to the optimization problems in Steps 1 and 2. As mentioned above, Step 3 is computationally inexpensive since it is static and can be computed offline (see Theorem 4). Furthermore, one can show that the optimization in Step 2 can be reduced to a formulation like the one in Step 1 in linear time. Thus, to assess the computational performance of the Myopic Tracking policy we will focus on analyzing the running time needed to solve the MLE problem in Step 1.

Recall that Step 1 is equivalent to a *weighted feedback arc set problem on tournaments* (see Proposition 3), which admits an integer programming (IP) formulation. In our numerical experiments, we solve this IP using an out-of-box solver as well as a heuristic based on its linear programming (LP) relaxation. We evaluate the running times of both methods and also record the relative optimality gaps of our heuristic in comparison to the LP relaxation of (15). Specifically, we generate a sequence of instances of the $\vec{w}$ data in (12) that arise as MTP iterates under the OA model and compute average running times and optimality gaps (relative to the LP relaxation) of these representative instances of (15). The out-of-the box solver we use is Gurobi 9.0.0 (win64, Python).[11] Both the Gurobi solver and our heuristic are run on an Intel Core i7-6700 CPU with a frequency of 3.40 GHz. Table 2 presents a summary of our numerical study and we refer the reader to Appendix J for more details on our heuristic and its implementation.

(K,p)	T_{IP} (sec)	T_H (sec)	Δ	(K,p)	T_{IP} (sec)	T_H (sec)	Δ
$(25,0.5)$	0.052	0.015	0%	$(25,0.9)$	0.059	0.014	0%
$(50,0.5)$	0.991	0.143	0%	$(50,0.9)$	0.837	0.143	0%
$(75,0.5)$	3.805	0.548	0%	$(75,0.9)$	4.576	0.695	0.001%
$(100,0.5)$	11.019	1.560	0%	$(100,0.9)$	13.692	1.878	0.028%
$(125,0.5)$	20.707	3.279	0%	$(125,0.9)$	30.576	3.808	0%
$(150,0.5)$	40.838	6.288	0%	$(150,0.9)$	60.025	6.383	0%

Table 2 Running time of the integer programming formulation T_{IP}, running time of our heuristic T_H, and relative optimality gap Δ of the heuristic for Step 1 of MTP.

The following are some of our main findings:

(i) The exact integer programming formulation can be solved within a reasonable amount of time, even when the number of items is relatively large. For example, even when $K = 100$ and $p = 0.9$, the average running time is within 15 seconds.

(ii) Our proposed heuristic is about an order of magnitude faster than the out-of-box solver. For example, even when $K = 100$ and $p = 0.9$, the average running time is less than 2 seconds.

(iii) In most cases, our heuristic achieves zero optimality gap with respect to the LP relaxation, and when the gap is nonzero, it is nonetheless small. In fact, our heuristic achieves zero optimality gap relative to the LP relaxation in 99.6% of the cases and the maximum gap we observe is 3.3%. This finding is consistent with those in related studies in earlier literature (see, e.g., Conitzer et al. 2006, Schalekamp and Zuylen 2009). Our interpretation of the effectiveness of the heuristic and the speed of our solver is that, while the *weighted feedback arc set problem on tournaments* is NP-hard in general, the actual choice data encountered by MTP concentrate on an "easier-to-solve" subclass of instances (at least under the OA model).

8.2. Sample Complexity of MTP

In this subsection, we evaluate the sample complexity (i.e., the average number of samples) of our proposed Myopic Tracking Policy and compare it to a number of alternative policies using two separate sets of numerical experiments. First, we consider three variants of the MTP policy that

[11] We note that we did not attempt to accelerate the implementation using techniques such as constraint generation, or problem-specific cutting plane methods as in Grötschel et al. (1984). Our main goal in this study is to test the performance of the integer programming formulation while not requiring special-purpose software.

employ different display sampling rules but similar stopping and recommendation rules. In our second set of experiments, we compare the MTP to three policies that have been proposed in the ranking and selection literature.

MTP-based Benchmark Policies: In our first set of numerical experiments we compare the sample complexity of the Myopic Tracking Policy against the following three variations:

1. The Full Display Policy, under which $S_t = [K]$ for all t.
2. The Pairwise Display Policy, under which S_t is randomized over the space of all pairs of items, i.e., $|S_t| = 2$. The randomization probabilities are given in equation (18) in Proposition 4 below. Using a similar analysis to one use for the Myopic Tracking Policy, one can show that this Pairwise Display Policy is worst-case asymptotically optimal in $\mathcal{M}_p^{\mathrm{OA}}$ if we restrict ourselves to policies that only use pairwise comparisons.
3. The Pair & Full Display Policy, under which S_t is randomized over the Top 2 (i.e., $\{\sigma^{-1}(1), \sigma^{-1}(2)\}$) and the full set $[K]$. The randomization probabilities are given in equation (19) below. Again, for this choice of randomization probabilities, one can show that this Pair & Full Display Policy is worst-case asymptotically optimal in $\mathcal{M}_p^{\mathrm{OA}}$ if we restrict ourselves to policies that only display the Top 2 versions or the full display set.

In what follows, we will use M, F, P, and PF to identify the Myopic Tracking policy, Full Display policy, Pairwise Display policy, and Pair & Full Display policy, respectively. Also, for a given policy $\pi \in \{\mathrm{M,F,P,\ PF}\}$, we let $\mathcal{S}^\pi \subseteq \mathcal{S}$ denote the set of display sets that are admissible under π, that is, $\mathcal{S}^M = \mathcal{S}$, $\mathcal{S}^P = \{S \subseteq [K] : |S| = 2\}$, $\mathcal{S}^F = \{[K]\}$, and $\mathcal{S}^{PF} = \{[2], [K]\}$.

Our next result justifies the specific choice of the randomized display strategies that we use in the implementation of the P and PF policies.

Proposition 4. *Let the randomization distributions* $\lambda_*^{\mathrm{P,OA}}(S) \in \Delta(\mathcal{S}^P)$, $\lambda_*^{\mathrm{PF,OA}}(S) \in \Delta(\mathcal{S}^{PF})$ *be given as:*

$$\lambda_*^{\mathrm{P,OA}}(S) := \begin{cases} \frac{1}{K-1} & \text{if } S = \{i, i+1\} \text{ for some } i \in \{1, \ldots, K-1\} \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

and

$$\lambda_*^{\mathrm{PF,OA}}(S) := \begin{cases} \frac{(\mathfrak{a}_3 - \mathfrak{a}_2)/\mathfrak{b}_K}{\mathfrak{a}_2/\mathfrak{b}_2 + (\mathfrak{a}_3 - \mathfrak{a}_2)/\mathfrak{b}_K} & \text{if } S = \{1, 2\} \\ \frac{\mathfrak{a}_2/\mathfrak{b}_2}{\mathfrak{a}_2/\mathfrak{b}_2 + (\mathfrak{a}_3 - \mathfrak{a}_2)/\mathfrak{b}_K} & \text{if } S = [K]. \end{cases} \tag{19}$$

where $\mathfrak{a}_n$ *and* $\mathfrak{b}_n$ *are defined in* (6)*. Then*

$$\lambda_*^{\mathrm{P,OA}} \in \operatorname*{arg\,max}_{\lambda \in \Delta(\mathcal{S}^P)} \min_{\bar{f} \in \overline{\mathcal{M}}_p(f_*^{\mathrm{OA}})} D_\lambda\left(f_*^{\mathrm{OA}} || \bar{f}\right) \quad \text{and} \quad \lambda_*^{\mathrm{PF,OA}} \in \operatorname*{arg\,max}_{\lambda \in \Delta(\mathcal{S}^{PF})} \min_{\bar{f} \in \overline{\mathcal{M}}_p(f_*^{\mathrm{OA}})} D_\lambda\left(f_*^{\mathrm{OA}} || \bar{f}\right).$$

The following proposition gives performance guarantees and theoretical justifications of our policies F, P and PF. The statement of Proposition 5 uses the shorthand notation $I^\pi = \max_{\lambda \in \Delta(\mathcal{S}^\pi)} \min_{\bar{f} \in \overline{\mathcal{M}}_p(f_*^{\mathrm{OA}})} D_\lambda\left(f_*^{\mathrm{OA}} || \bar{f}\right)$.

PROPOSITION 5. *With the stopping threshold $\beta = C_1 + \log(1/\delta)$ as in Theorem 7, all of the policies $\{\mathrm{F}, \mathrm{P}, \mathrm{PF}\}$ are δ-accurate for every $\delta \in (0,1)$. Their sample complexities are such that for all $f \in \mathcal{M}_p$,*

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_f^F[\tau]}{\log\left(\frac{1}{\delta}\right)} \leq \frac{1}{I^F}, \quad \limsup_{\delta \downarrow 0} \frac{\mathbb{E}_f^P[\tau]}{\log\left(\frac{1}{\delta}\right)} \leq \frac{1}{I^P}, \quad \textit{and} \quad \limsup_{\delta \downarrow 0} \frac{\mathbb{E}_f^{PF}[\tau]}{\log\left(\frac{1}{\delta}\right)} \leq \frac{1}{I^{PF}}, \tag{20}$$

where

$$I^F = (1-p)\log\left(\tfrac{1}{p}\right)\frac{1}{(1-p^K)/(1-p)}, \qquad I^P = (1-p)\log\left(\tfrac{1}{p}\right)\frac{1}{(K-1)(1+p)} \quad \textit{and}$$

$$I^{PF} = (1-p)\log\left(\tfrac{1}{p}\right)\frac{1+2p}{(1-p^K)/(1-p) + 2p(1+p)}.$$

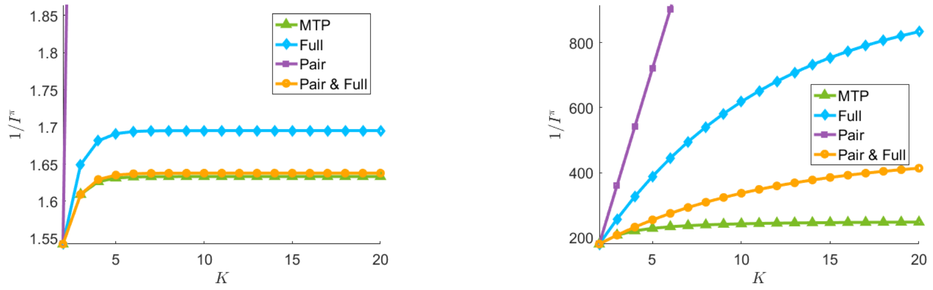
Moreover, if $f \in \mathcal{M}_p^{\mathrm{OA}}$ then

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}_f^F[\tau]}{\log\left(\frac{1}{\delta}\right)} = \frac{1}{I^F} \quad \textit{and} \quad \lim_{\delta \downarrow 0} \frac{\mathbb{E}_f^P[\tau]}{\log\left(\frac{1}{\delta}\right)} = \frac{1}{I^P}. \tag{21}$$

As a result, F *and* P *are worst-case asymptotically optimal within the classes of full and pairwise display δ-accurate policies, respectively. That is, if we let $\mathcal{A}^F$ (resp. $\mathcal{A}^P$) be the space of accurate policies such that $S_t \in \mathcal{S}^F$ (resp. $S_t \in \mathcal{S}^P$), we have*

$$\mathrm{F} \in \arg\min_{\pi \in \mathcal{A}^F} \sup_{f \in \mathcal{M}_p} \limsup_{\delta \downarrow 0} \frac{\mathbb{E}_f^\pi[\tau]}{\log(1/\delta)} \quad \textit{and} \quad \mathrm{P} \in \arg\min_{\pi \in \mathcal{A}^P} \sup_{f \in \mathcal{M}_p} \limsup_{\delta \downarrow 0} \frac{\mathbb{E}_f^\pi[\tau]}{\log(1/\delta)}. \tag{22}$$

We use Proposition 5 to conduct a sensitivity analysis of the policies $\pi \in \{\mathrm{M,F,P,\ PF}\}$ on p and K when δ is sufficiently small. Figure 4 depicts the values of the $1/I^\pi$ for different values of p and K.



(a) $p = 0.3$

(b) $p = 0.9$

Figure 4 I^π as function of K for $p = 0.3$ and $p = 0.9$.

We can see from the figure that the Myopic Tracking Policy offers a significant advantage over the other policies. In the instance we looked at, the pairwise policy performed particularly poorly, whereas pairwise and full remained closest in performance to MTP. This is consistent with earlier findings as in Figure 3, where we saw MTP emphasize pairwise and full display sets in its randomized display set policy.

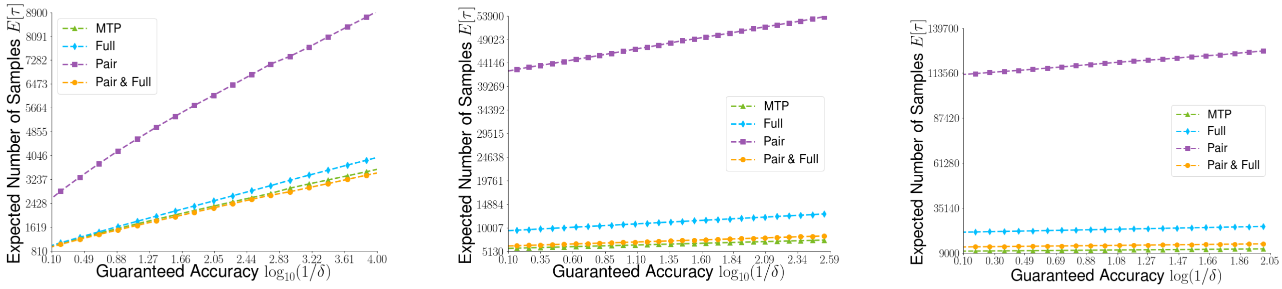
Let us now conduct a set of numerical simulations to compare the value of $\mathbb{E}^{\pi}_{f_*}[\tau]$ for $\pi \in \{\text{M, F, P, PF}\}$ and to investigate the sensitivity of the MTP sample complexity to δ. In our simulation study, we fix $p = 0.9$, and set $K \in \{4, 10, 15\}$ with appropriate values of β.

We take the underlying choice model to belong to $\mathcal{M}^{\text{OA}}_p$, because that is the worst-case model given every $\mathcal{M}_p$.

For each problem instance (K, p), we evaluate two metrics simultaneously:

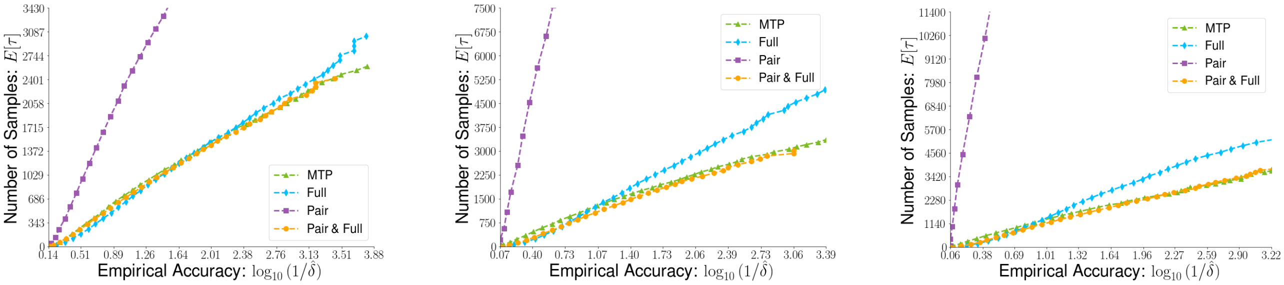
- The sample complexity as a function of δ for the learning algorithm to be δ-accurate. This metric is the closest to what we theoretically studied earlier. While many values of β are appropriate (i.e., guarantee low error probability and asymptotically optimal sample complexity), we select $\beta = C_1 + \log\left(\frac{1}{\delta}\right)$, where $C_1 = \log\big((K-1)(K-1)!\big)$. Note that the current value of β is only an upper bound of what is needed to guarantee δ-accuracy. With that said, it is the smallest value we know in our proof. Moreover, the resulting loss of performance by picking a conservative β is negligible asymptotically as $\delta \downarrow 0$;
- The sample complexity as a function of the empirical error probability $\hat{\delta}$, as we vary different levels of β. Here the empirical error probability is defined as the fraction of instances in which the algorithm terminates with an item different from $\sigma^{-1}_{f_*}(1)$ under the preference f_*. This metric characterizes where a learning algorithm stands in the trade-off between speed (i.e., being able to stop early) and accuracy (i.e., achieving low error probability) under the specific preference model f_*.

The performance of the four policies {M, F, P, PF} under the OA model is illustrated in Figures 5 and 6. Our computational experiments suggest that policy P is significantly outperformed by policy F which, in turn, is significantly outperformed by policies PF and M. Furthermore, the performance gap among these policies increases with K and $1/\delta$, which is consistent with our theoretical analysis. It is interesting to note that the empirical errors of PF and M are comparable,



(a) $K = 4$ (b) $K = 10$ (c) $K = 15$

Figure 5 Sample complexity vs. theoretically guaranteed error probability (log scale).



(a) $K = 4$

(b) $K = 10$

(c) $K = 15$

Figure 6 Sample complexity vs. empirical error probability (log scale)

further suggesting that the PF policy might be the right policy to use in practice as it offers a good compromise between performance and implementation simplicity. This echoes our previous discussion on the need for an optimal policy to balance the trade-off between accuracy and coverage. Our numerical results suggest that only displaying pairs and the full set is close to an optimal policy, if one selects the randomization probabilities carefully (that is, according to equation (19)).

Benchmarks from the Ranking and Selection Literature: We conclude our numerical experiments by comparing the performance of the MTP policy to three alternative policies that have been proposed in the literature on ranking and selection under noisy pairwise comparisons. Specifically, we consider policies AR and AR2 proposed by Heckel et al. (2019) and policy PLPAC proposed by Szörényi et al. (2015). Since these policies were developed in the context of pairwise comparisons, we also include policy P in our numerical experiments. Also, the parameters of the different algorithms were selected so that all policies are δ-accurate[12] with $\delta = 0.01$ for the different problem instances (K, p) that we consider.

(K,p)	M $(\times 10^4)$	P $(\times 10^4)$	AR $(\times 10^4)$	AR2 $(\times 10^4)$	PLPAC $(\times 10^4)$
$(4,0.5)$	0.0048	0.014	0.235	0.162	0.261
$(10,0.5)$	0.0134	0.125	2.49	1.77	9.95
$(15,0.5)$	0.0207	0.302	6.29	4.57	1.68

(K,p)	M $(\times 10^4)$	P $(\times 10^4)$	AR $(\times 10^4)$	AR2 $(\times 10^4)$	PLPAC $(\times 10^4)$
$(4,0.9)$	0.231	0.613	9.67	6.73	14.4
$(10,0.9)$	0.712	5.13	101	73.2	53.2
$(15,0.9)$	1.13	12.5	259	188	88.9

Table 3 Sample complexity comparisons for $\delta = 0.01$.

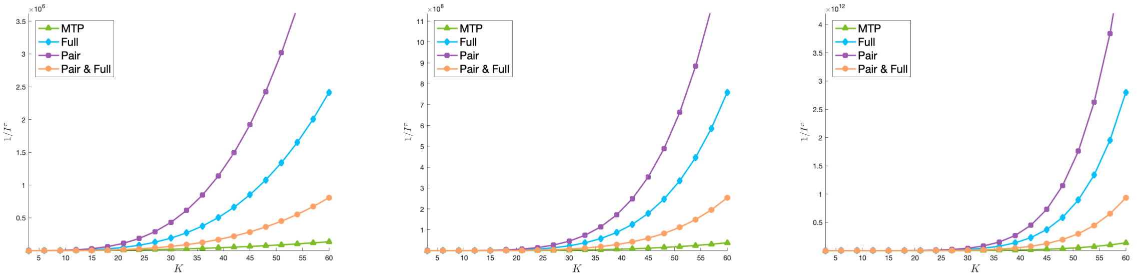
Table 3 presents a summary of our numerical study. We find that policy M uses on average an order of magnitude fewer samples than policy P which, in turn, uses an order of magnitude

[12] PLPAC need the additional assumption that the properties of the Bradley-Terry-Luce model holds.

fewer samples than the AR, AR2, and PLPAC policies. These results show promise for the MTP methodology, but we also need to take them 'with a grain of salt'. The reason is that policies AR, AR2, and PLPAC were developed without imposing any separability requirement on the underlying choice model and so they do not explicitly take advantage –as the M and P policies do– of the fact that preferences are ranking based or that they belong to the set $\mathcal{M}_p$. Nevertheless, our results shed some light on the question of how much one can gain from incorporating additional structure (e.g., parametric and/or separability assumptions) into the model. While the work of Heckel et al. (2019) reveals that imposing specific type of parametric assumptions (e.g., consumers' choice preferences belong to the popular class of Bradley-Terry-Luce models) offers limited benefits for stochastic comparisons, our numerical results suggest that imposing some mild structure (as in Definition 1) goes a long way in reducing sample complexity.

Asymptotic Performance as $p \uparrow 1$ and $K \uparrow \infty$: Motivated by our discussion in Remark 2, let us conclude this section investigating the sample complexity of the Myopic Tracking Policy as $p \uparrow 1$ and $K \uparrow \infty$ jointly. As noted in Remark 2, under this type of asymptotic regime, the set $\mathcal{M}_p$ of p-separable preferences could include Luce-type preferences whose attraction scores are not required to decay exponentially fast as the number of products grow large.

Figure 7 depicts the value of $1/I^\pi$ as a function of K for four policies $\pi \in \{\mathrm{M}, \mathrm{F}, \mathrm{P}, \mathrm{PF}\}$. Each panel consider a different regime describing the dependence of p on K: (a) $p = 1 - 1/\log K$; (b) $p = 1 - 1/K$; and (c) $p = 1 - 1/K^2$. We find that when $K \uparrow \infty$ and $p \uparrow 1$ jointly, the values of $1/I^\pi$ under Full, Pairwise, and Pair & Full display policies perform arbitrarily bad compared to that of MTP. This finding is robust across the three regimes.



(a) $p = 1 - 1/\log K$

(b) $p = 1 - 1/K$

(c) $p = 1 - 1/K^2$

Figure 7 I^π as a function of K for $p = 1 - 1/\log K$, $p = 1 - 1/K$, and $p = 1 - 1/K^2$ respectively.

To further support to these numerical experiments, the following proposition compares the sample complexities of Policies P and F to MTP as $p \uparrow 1$ and $K \uparrow \infty$ jointly.

Proposition 6. *For $f \in \mathcal{M}_p^{\mathrm{OA}}$ and for any $p \in [0,1)$ and $K \in \mathbb{N}$*

$$\lim_{\delta\downarrow 0} \frac{\mathbb{E}_f^F[\tau]}{\mathbb{E}_f^M[\tau]} \geq \frac{K(1+p+\cdots+p^{K-1})}{K+2p(K-1)} \qquad \textit{and} \qquad \lim_{\delta\downarrow 0} \frac{\mathbb{E}_f^P[\tau]}{\mathbb{E}_f^M[\tau]} \geq \frac{K\,(K-1)(1+p)}{K+2p(K-1)}.$$

It follows that

$$\lim_{p\uparrow 1, K\uparrow\infty} \; \lim_{\delta\downarrow 0} \frac{\mathbb{E}_f^F[\tau]}{\mathbb{E}_f^M[\tau]} = \lim_{p\uparrow 1, K\uparrow\infty} \; \lim_{\delta\downarrow 0} \frac{\mathbb{E}_f^P[\tau]}{\mathbb{E}_f^M[\tau]} = \infty.$$

According to this proposition, the policies P and F can perform "arbitrarily bad" compared to MTP in the asymptotic regimes when $p \uparrow 1$, $K \uparrow \infty$, and $\delta \downarrow 0$. Notice that this is independent of the specific regime under consideration, as long as $p \uparrow 1$ as $K \uparrow \infty$.

Finally, we can get a similar result if we compare the asymptotic sample complexity of MTP to the AR and AR2 policies discussed above. To this end, note that from Proposition 5, the sample complexity of policy P satisfies

$$\lim_{\delta\downarrow 0} \frac{\mathbb{E}_f^P[\tau]}{\log(1/\delta)} = \frac{(1+p)(K-1)}{(1-p)\log(1/p)} \quad \text{for any } f \in \mathcal{M}_p^{\mathrm{OA}}.$$

Also, from Theorem 1b in Heckel et al. 2019 and Theorem 1 in Jamieson et al. 2015 we have that

$$\lim_{\delta\downarrow 0} \frac{\mathbb{E}_f^\pi[\tau]}{\log(1/\delta)} \geq \frac{(K-1)^2(1+p)^2}{(1-p)^2} \sum_{i=2}^{K} \frac{1}{(i-1)^2}, \qquad \text{for any } f \in \mathcal{M}_p^{\mathrm{OA}} \quad \text{and } \pi \in \{\text{AR,AR2}\}.$$

So for a given p, policy P has an asymptotic sample complexity $O(K)$ while the AR and AR2 have asymptotic sample complexity $\Omega(K^2)$.[13] If we take $p \uparrow 1$ and $K \uparrow \infty$ jointly, we have for any $f \in \mathcal{M}_p^{\mathrm{OA}}$ and $\pi \in \{\text{AR,AR2}\}$,

$$\lim_{K\uparrow\infty,\; p\uparrow 1} \; \lim_{\delta\downarrow 0} \frac{\mathbb{E}_f^\pi[\tau]}{\mathbb{E}_f^P[\tau]} \geq \lim_{K\uparrow\infty,\; p\uparrow 1} \frac{(K-1)(1+p)\log(1/p)}{(1-p)} \sum_{i=2}^{K} \frac{1}{(i-1)^2} = \infty. \tag{23}$$

That is, both AR and AR2 can perform "arbitrarily bad" compared to Policy P and, hence by Proposition 6, to the Myopic Tracking Policy). This is also independent of the asymptotic regime $p \uparrow 1$ and $K \uparrow \infty$.

9. Concluding Remarks and Further Directions

In this paper, we have studied a class of ranking and selection problems faced by a company that wants to identify the most preferred product out of a finite set of alternatives when consumer preferences are *a priori* unknown. Specifically, we have assumed that the only information available

[13] Note that this observation does not contradict the results in Heckel et al. (2019) and Jamieson et al. (2015) because they need to identify a δ-accurate algorithm over preference models where the consistency assumption (A-3) and the separability assumption (A-4) may be violated. Our finding does not violate Theorem 2a in Heckel et al. (2019) either because under their parametric structure, the lower bound for the OA model is vacuous: Φ is a step function, and therefore, $\phi_{\min} = 0$, $\phi_{\max} = \infty$, and $c_{\text{par}} = 0$ in their notations.

is that consumer preferences belong to a class $\mathcal{M}_p$ that satisfies two key properties: (i) choice probabilities are consistent with some unknown true ranking of the alternatives, and (ii) they satisfy a mild separability condition under which no two products are equally preferred (i.e., consumers have strict preferences over the different alternatives). To learn the unknown ranking over the products, we have assumed that the company is able to sample consumer preferences by sequentially showing different subsets of products to different consumers and asking them to report their top preference within the display set they were offered. In this setting, we have formulated the problem as one of designing a display policy that minimizes the expected number of samples needed to identify the top-ranked product for a given error probability $\delta \in (0,1)$.

Because of the minimal assumptions imposed on consumer preferences, we proposed a *robust learning* methodology to derive a display policy, our *Myopic Tracking Policy* (MTP), that is worst-case asymptotically optimal as $\delta \downarrow 0$. This means that for any other policy π there exists an error probability δ and a consumer's preference in $\mathcal{M}_p$ under which the expected number of samples needed to identify the top-ranked version using MTP is less than or equal to the expected number of samples needed by policy π. Besides this theoretical performance guarantee, the Myopic Tracking policy also shows good non-asymptotic numerical performance. Through a set of computational experiments, we showed that the Myopic Tracking policy has consistently better sample performance (i.e., learns faster for a given level of accuracy) when compared to various alternative policies. Our numerical experiments also show that the Full & Pair Display policy –a variation of the MTP policy in which only pairs of the full set are displayed– offers a good compromise between performance and simplicity, which makes it particularly appealing from a practical standpoint.

In terms of future work, we envision a few directions in which our results can be extended. First, in the context of the top-ranked identification problem that we have considered in this paper, one could explore the possibility of relaxing some of the requirements that we have imposed on the set $\mathcal{M}_p$ to consider a larger set $\widetilde{\mathcal{M}}_p$ of admissible preferences. In particular, rather than requiring that consumers have strict preferences over the entire menu of products, we could only require strict preference for the top-ranked product, allowing for indifference among the rest of the products. This can be accomplished by replacing conditions (A-3) and (A-4) in Definition 1 by the following weaker condition:

(A-5) For every $f \in \widetilde{\mathcal{M}}_p$ there exists $X_f \in [K]$ such that $p\, f(X_f|S) \geq f(X'|S)$ for all $S \ni X_f$.

Extending our robust framework to this larger set $\widetilde{\mathcal{M}}_p$ of consumer preference requires a number of changes in our methodology. We anticipate that the most challenging adjustment would be to find the set of hardest-to-learn preferences and adapting our analysis accordingly. We believe, however, that our proposed Myopic Tracking policy will still perform well in this more general setting.

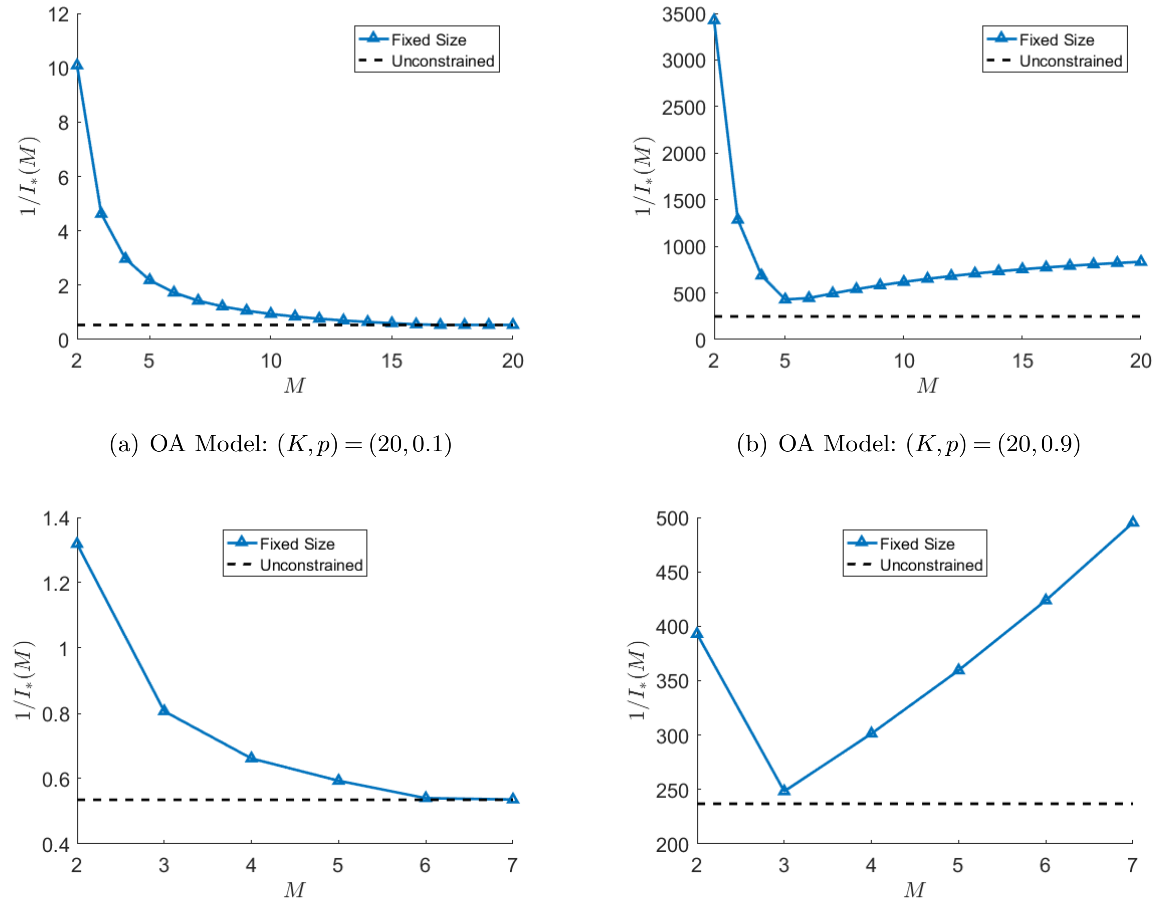
Although formal support of this claim is beyond the scope of this paper, we have conducted a set of numerical experiments using a real data set for which the strict ranking assumption does not hold, and yet the Myopic Tracking policy still performs well when compared to other benchmark policies. The data comes from a survey conducted at the AGH University of Science and Technology in which students were asked to provide a rank ordering over a set of courses with no missing elements (further details of the dataset and the numerical experiments are provided in Appendix K).

A second direction that we believe is worth exploring is to extend our results on instance-specific sample optimality. The analysis in this paper gives us a good understanding of how to achieve instance-specific optimality if we restrict the set of admissible preferences to any arbitrarily finite subset of $\mathcal{M}_p$ (Theorem 2), as well as how to achieve worst-case optimality for the whole of $\mathcal{M}_p$ (Theorem 5). We conjecture that a Max-Min-type problem is still central in developing an asymptotically optimal algorithm.

Another direction in which our paper can be extended relates to how to push our analysis for moderately-sized δ and to derive a policy that is "higher-order" optimal than the Myopic Tracking Policy. In fact, Theorem 6 implies that under the MTP policy, $\mathbb{E}_f^{\tilde{\pi}}[\tau] \le \frac{\log(1/\delta)}{I_*^{\text{OA}}} + o(\log\frac{1}{\delta})$ for all $f \in \mathcal{M}_p$. We conjecture that there exists an algorithm $\tilde{\pi}$ such that $\mathbb{E}_f^{\tilde{\pi}}[\tau] \le \frac{\log(1/\delta)}{I_*^{\text{OA}}} + O(1)$ for all $f \in \mathcal{M}_p$, where the improvement is in the residual term from $o(\log\frac{1}{\delta})$ to $O(1)$ as a function of δ.

Additionally, using the current framework of analysis, there is a potential that the Myopic Tracking Policy can be generalized to a broader class of problem formulations. For example, we could consider (i) a constrained set $\mathcal{S} \subseteq \mathcal{P}([K])$ of possible display sets; (ii) other problem objectives such as top-k selection, or the full ranking identification problem; (iii) a general class of probabilistic choice model $f(\cdot|S)$ that goes beyond p-separability; or even (iv) a more general feedback mechanism such as asking consumers to provide full rankings rather than single choices. Of course, with these extensions, the strategies (i.e. the randomization distribution) will change as the Max-Min problem changes. So will the MLE solution change if we change the probabilistic model $f(\cdot|S)$. We are interested in understanding the structural properties of MTP under these different formulations. For instance, we conjecture that the randomization distribution, as a result of the new Max-Min problem, is sparse in general.

As mentioned above, our computational experiments in Section 8 revealed that the simple Full & Pair Display policy can achieve good performance. This raises the question of whether there are other simple policies that could also have good performance. This is a particularly important issue in many practical settings in which companies are restricted, or must commit, to display sets of a given fixed cardinality (see Vinayak and Hassibi 2016 for a discussion comparing display sets of cardinality two and three). It would be interesting to extend our methodology to identify an optimal policy within the class of strategies that use display sets of a given size. To illustrate this point, suppose the company is restricted to offer display sets of fixed size M and let $I_*(M)$ denote Chernoff's information measure subject to this additional cardinality constraint. Figure 8 depicts the value of $1/I_*(M)$ (solid line) and the unconstrained value of $1/I_*$ (dashed line) as a function of M for different values of K and p and two preference models: OA model (top panels) and Mallows model (bottom panels).



(a) OA Model: $(K,p) = (20, 0.1)$

(b) OA Model: $(K,p) = (20, 0.9)$

(c) Mallows Model: $(K,p) = (7, 0.1)$

(d) Mallows Model: $(K,p) = (7, 0.9)$

Figure 8 Chernoff's inverse information measure $1/I_*(M)$ as a function of the display set cardinality M for different values of K and p and two preference models: OA model (top panels) and Mallows model (bottom panel).

These preliminary results suggest that, for low values of p, a display policy that maximizes the cardinality of the display set could be optimal among those policies with fixed cardinality. Furthermore, Chernoff's inverse information measure for the full display policy is almost identical to the unconstrained value. On the other hand, when p is large is not true that a policy that maximizes the cardinality is better. In the example above, for the OA model with $(K,p) = (20, 0.9)$

the optimal fixed size display set is $M^*=5$ and for the Mallows model with $(K=7, p=0.9)$ the optimal fixed size display set is $M^*=3$. In general, it would be interesting to understand how M^* changes as a function of K and p as well as what is the optimality of gap (measured by the difference $1/I_*(M^*)-1/I_*$) resulting from using a fixed size display strategy.

Finally, we also see room for extending the optimization and computational methods used in the implementation of the MTP policy. For instance, one can think of developing more specialized methods that take full advantage of the structure of the problem to speed up the solution time of the MLE problem in Steps 1 and 2 of Algorithm 1. Similarly, one could consider variations of the MTP policy in which the MLE problem is not solved in every iteration but at a less frequent rate (e.g., at a monotonically increasing sequence of (possibly random) time epochs $1 \le \tau_1 < \tau_2 < \cdots$). In theory, by carefully selecting the values of $\{\tau_i\}$, one could increase the speed of the algorithm without a significant decay in performance.

References

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2017) Thompson sampling for the MNL-bandit. *COLT*, 76–78.

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2019) MNL-bandit: A dynamic learning approach to assortment selection. *Operations Research* 67(5):1453–1485.

Ailon N (2012) An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *J. Mach. Learn. Res.* 13(Jan):137–164.

Ailon N, Charikar M, Newman A (2005) Aggregating inconsistent information: Ranking and clustering. *STOC*, 684–693.

Ailon N, Charikar M, Newman A (2008) Aggregating inconsistent information: ranking and clustering. *Journal of the ACM* 55(5):23.

Ali A, Meila M (2012) Experiments with Kemeny ranking: What works when? *Math. Social Sci.* 64(1):28–40.

Alon N (2006) Ranking tournaments. *SIAM J. Discrete Math.* 20(1):137–142.

Araman V, Caldentey R (2016) Crowdvoting the timing of new product introduction. *Available at SSRN: https://ssrn.com/abstract=2723515* .

Audibert JY, Bubeck S (2010) Best arm identification in multi-armed bandits. *COLT*, 41–53.

Betabrand (2018) URL https://www.betabrand.com/.

Brabham CD (2010) Moving the crowd at Threadless. *Inf. Commun. Soc.* 13:1122–1145.

Braverman M, Mossel E (2008) Noisy sorting without resampling. *SODA*, 268–276.

Braverman M, Mossel E (2009) Sorting from noisy information. *arXiv preprint arXiv:0910.1191* .

Bubeck S, Munos R, Stoltz G (2011) Pure exploration in finitely-armed and continuous-armed bandits. *Theor. Comput. Sci.* 412(19):1832–1852.

Caragiannis I, Procaccia AD, Shah N (2013) When do noisy votes reveal the truth? *EC*, 143–160.

Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. *Manag. Sci.* 53(2):276–292.

Charbit P, Thomassé S, Yeo A (2007) The minimum feedback arc set problem is NP-hard for tournaments. *Comb. Probab. Comput.* 16(1):1–4.

Charon I, Hudry O (2010) An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Ann. Oper. Res.* 175(1):107–158.

Chen X, Li Y, Mao J (2018) A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. *SODA*, 2504–2522.

Chen X, Wang Y (2017) A note on tight lower bound for MNL-bandit assortment selection models. *arXiv preprint arXiv:1709.06109* .

Chernoff H (1959) Sequential design of experiments. *Ann. of Math. Stat.* 30(3):755–770.

Chernoff H (1972) *Sequential Analysis and Optimal Design* (SIAM).

Chung F, Lu L (2006) Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics* 3(1):79–127.

Conitzer V, Davenport A, Kalagnanam J (2006) Improved bounds for computing Kemeny rankings. *AAAI*, 620–626.

Dantzig G (1963) *Linear Programming and Extensions* (Princeton, NJ: Princeton Univ. Press).

Davenport A, Kalagnanam J (2004) A computational study of the Kemeny rule for preference aggregation. *AAAI*, 697–702.

Désir A, Goyal V, Jagabathula S, Segev D (2018) Mallows-smoothed distribution over rankings approach for modeling choice. *Available at SSRN 3172997* .

Draglia V, Tartakovsky AG, Veeravalli VV (1999) Multihypothesis sequential probability ratio tests. I. Asymptotic optimality. *IEEE Trans. Inform. Theory* 45(7):2448–2461.

Falahatgar M, Orlitsky A, Pichapati V, Suresh AT (2017) Maximum selection and ranking under noisy comparisons. *ICML*, 1088–1096.

Fomin FV, Lokshtanov D, Raman V, Saurabh S (2010) Fast local search algorithm for weighted feedback arc set in tournaments. *AAAI*, 65–70.

Gabillon V, Ghavamzadeh M, Lazaric A (2012) Best arm identification: A unified approach to fixed budget and fixed confidence. *NIPS*, 3212–3220.

Garivier A, Kaufmann E (2016) Optimal best arm identification with fixed confidence. *COLT*, 998–1027.

Grötschel M, Jünger M, Reinelt G (1984) A cutting plane algorithm for the linear ordering problem. *Oper. Res.* 32(6):1195–1220.

Heckel R, Shah NB, Ramchandran K, Wainwright MJ, et al. (2019) Active ranking from pairwise comparisons and when parametric assumptions do not help. *Ann. of Stat.* 47(6):3099–3126.

Huang Y, Singh VP, Srinivasan K (2014) Crowdsourcing new product ideas under consumer learning. *Manag. Sci.* 60.

Jamieson KG, Katariya S, Deshpande A, Nowak RD (2015) Sparse dueling bandits. *AISTATS*, 416–424.

Jiang X, Lim LH, Yao Y, Ye Y (2011) Statistical ranking and combinatorial Hodge theory. *Math. Program.* 127(1):203–244.

Kaufmann E, Cappé O, Garivier A (2016) On the complexity of best arm identification in multi-armed bandit models. *J. Mach. Learn. Res.* 17(1):1–42.

Kenyon-Mathieu C, Schudy W (2007) How to rank with few errors. *STOC*, 95–103 (ACM).

King A, Lakhani KR (2013) Using open innovation to identify the best ideas. *MIT Sloan Man. Rev.* 55(1):41.

Lego (2018) URL https://ideas.lego.com/.

Levin J, Nalebuff B (1995) An introduction to vote-counting schemes. *J. Econ. Perspect.* 9(1):3–26.

Li X, Liu J, Ying Z (2014) Generalized sequential probability ratio test for separate families of hypotheses. *Sequential Analysis* 33(4):539–563.

Marinesi S, Girotra K (2012) Information acquisition through customer voting systems. *Available at SSRN: https://ssrn.com/abstract=2191940* ISSN 1556-5068.

Mitchell JE, Borchers B (1996) Solving real-world linear ordering problems using a primal-dual interior point cutting plane method. *Ann. Oper. Res.* 62(1):253–276.

Naghshvar M, Javidi T, et al. (2013) Active sequential hypothesis testing. *Ann. of Stat.* 41(6):2703–2738.

Pee LG (2016) Customer co-creation in B2C e-commerce: Does it lead to better new products? *Elec. Commerce Res.* 16(2):217–243.

PREFLIB (2019) URL http://www.preflib.org/.

Raykar CV, Yu S, Zhao HL, Valadez HG, Florin C, Bogoni L, Moy L (2010) Learning from crowds. *J. Mach. Learn. Res.* 11:1297–1322, ISSN 1532-4435.

Rusmevichientong P, Shen ZJM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Oper. Res.* 58(6):1666–1680.

Russo D (2016) Simple Bayesian algorithms for best arm identification. *COLT*, 1417–1418.

Sauré D, Zeevi A (2013a) Optimal dynamic assortment planning with demand learning. *Manuf. Serv. Oper. Manag.* 15(3):387–404.

Sauré D, Zeevi A (2013b) Optimal dynamic assortment planning with demand learning. *Manuf. Serv. Oper. Manag.* 15(3):387–404.

Schalekamp F, Zuylen Av (2009) Rank aggregation: Together we're strong. *ALENEX*, 38–51 (SIAM).

Schneider J, Hall J (2011) Why most product launches fail. *Harvard Bus. Rev.* (April):21–23.

Shah NB, Wainwright MJ (2017) Simple, robust and optimal ranking from pairwise comparisons. *J. Mach. Learn. Res.* 18(1):7246–7283.

Szörényi B, Busa-Fekete R, Paul A, Hüllermeier E (2015) Online rank elicitation for Plackett-Luce: A dueling bandits approach. *NeurIPS*, 604–612.

Ulu C, Honhon D, Alptekinoglu A (2012) Learning consumer tastes through dynamic assortments. *Oper. Res.* 60(4):833–849.

Van Zuylen A, Williamson DP (2009) Deterministic pivoting algorithms for constrained ranking and clustering problems. *Math. Oper. Res.* 34(3):594–620.

Vinayak RK, Hassibi B (2016) Crowdsourced clustering: Querying edges vs triangles. *Advances in Neural Information Processing Systems*, 1316–1324.

Wald A (1973) *Sequential Analysis* (Courier Corporation).

Wauthier FL, Jordan MI, Jojic N (2013) Efficient ranking from pairwise comparisons. *ICML*, III–109–117.

Young HP (1988) Condorcet's theory of voting. *Am. Political Sci. Rev.* 82(4):1231–1244.

Appendix A: Proof of Proposition 1

Let f_{M} be a consumer preference induced by a Mallows model with modal ranking σ_0 and concentration parameter $\theta > 0$. That is,

$$f_{\mathrm{M}}(x|S) = \sum_{\sigma\in\Sigma:\ \sigma(x)<\sigma(i), \forall i\in S\setminus\{x\}} \mathscr{P}(\sigma) = \sum_{\sigma\in\Sigma:\ \sigma(x)<\sigma(i), \forall i\in S\setminus\{x\}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}.$$

Here recall that $d(\sigma,\sigma_0) = \sum_{i<j} \mathbb{I}\{\sigma(i) > \sigma(j)\}$ is the Kendall-Tau distance between ranking σ and the modal ranking σ_0 and $\Psi(\theta)$ is a normalization parameter that only depends on θ. Without loss of generality, suppose that the modal ranking σ_0 is the identity ranking σ_*.

To verify that $f_{\mathrm{M}} \in \mathcal{M}_p$ with $p = e^{-\theta}$, notice that it is straightforward to verify (A-1) and (A-2), and (A-3) is implied by (A-4) because $p < 1$. Hence it suffices to verify that (A-4) holds. That is, it suffices to show that for every display set S and $k, j \in S$ such that $k < j$, the preference f_{M} satisfies

$$\frac{f_{\mathrm{M}}(k|S)}{f_{\mathrm{M}}(j|S)} \geq e^{\theta}.$$

Note that here we are implicitly taking $\sigma_{\mathrm{M}} = \sigma_*$ in (A-4).

Throughout the proof. let us fix a display set S and single out two elements k, j in S so that $S = \{k, j\} \cup S_0$ for some (possibly empty) set S_0. Without loss of generality, suppose that $k < j$ and j is the immediate successor of k in S. That is, $\{k+1, \ldots, j-1\} \cap S = \emptyset$. We also let $B_x := \{\sigma : \sigma(x) < \sigma(i),\ \text{for all } i \in S_0\}$ for every $x \in S \setminus S_0$.

Case 1. k and j are adjacent, i.e., $k = j-1$. On a high level, we can pair the rankings that rank $j-1$ at the top among display set S with those that rank j at the stop among display set S, so that each of the former has exactly one less pairwise inconsistency than each of the latter. Consider the following bijection $\phi : \Sigma \to \Sigma$ that swaps the ranking of j and $j-1$:

$$\phi(\sigma)(i) = \begin{cases} \sigma(j) & \text{if } i = j-1 \\ \sigma(j-1) & \text{if } i = j \\ \sigma(i) & \text{otherwise.} \end{cases}$$

We refer the reader to the proof of Theorem EC.1 in Désir et al. (2018) for the following property of ϕ:

$$d(\phi(\sigma), \sigma_0) = d(\sigma, \sigma_0) + 1, \qquad \text{for every } \sigma \in A := \{\sigma : \sigma(j-1) < \sigma(j)\};$$

$$d(\phi(\sigma), \sigma_0) = d(\sigma, \sigma_0) - 1, \qquad \text{for every } \sigma \in A^c := \{\sigma : \sigma(j) < \sigma(j-1)\}.$$

That is, the swapping mapping increases/decreases the total number of pairwise inconsistencies by exactly one. Recall $S = \{j-1, j\} \cup S_0$, $B_{j-1} = \{\sigma : \sigma(j-1) < \sigma(i),\ \text{for all } i \in S_0\}$ and $B_j = \{\sigma : \sigma(j) < \sigma(i),\ \text{for all } i \in S_0\}$. Hence

$$\{\sigma : \sigma(j-1) < \sigma(i),\ \forall i \in S \setminus \{j-1\}\} = \{\sigma : \sigma(j-1) < \sigma(j)\} \cap \{\sigma : \sigma(j-1) < \sigma(i),\ \forall i \in S_0\} = A \cap B_{j-1}$$
$$\{\sigma : \sigma(j) < \sigma(i),\ \forall i \in S \setminus \{j\}\} = \{\sigma : \sigma(j) < \sigma(j-1)\} \cap \{\sigma : \sigma(j) < \sigma(i),\ \forall i \in S_0\} = A^c \cap B_j.$$

As a consequence,

$$f_{\mathrm{M}}(j-1|S)=\sum_{\sigma\in B_{j-1}\cap A}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}=\sum_{\sigma\in B_{j-1}\cap A}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}\overset{(a)}{=}\sum_{\sigma\in B_j\cap A^c}\frac{e^{-\theta d(\phi(\sigma),\sigma_0)}}{\Psi(\theta)}=e^{\theta}\sum_{\sigma\in B_j\cap A^c}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}=e^{\theta}f_{\mathrm{M}}(j|S).$$

That is, $\frac{f_{\mathrm{M}}(k|S)}{f_{\mathrm{M}}(j|S)}=e^{\theta}$. In the derivation above, part (a) uses the fact that the bijection ϕ restricted to the set $B_j\cup A^c$ is also a bijection from $B_j\cup A^c$ to $B_{j-1}\cup A$.

Case 2. k and j are not adjacent, i.e., $k<j-1$. On a high level, it means that k and j are further away, and hence the choice probabilities should be more concentrated on k over j. To formally show this, let us a consider a local adjustment to the set S. Let $\tilde{S}:=S\setminus\{j\}\cup\{j-1\}=\{k,j-1\}\cup S_0$. That is, $\tilde{S}$ is obtained by replacing version j by version $j-1$ in the display set S. It suffices to show that

$$f_{\mathrm{M}}(k|S)\geq f_{\mathrm{M}}(k|\tilde{S}) \tag{24}$$

$$f_{\mathrm{M}}(j|S)\leq f_{\mathrm{M}}(j-1|\tilde{S}), \tag{25}$$

because in this case, we can further replace version $j-1$ by version $j-2$, and so on. By induction,

$$\begin{aligned}\frac{f_{\mathrm{M}}(k|S)}{f_{\mathrm{M}}(j|S)}=&\frac{f_{\mathrm{M}}(k|\{k,j\}\cup S_0)}{f_{\mathrm{M}}(j|\{k,j\}\cup S_0)}\\ \geq&\frac{f_{\mathrm{M}}(k|\{k,j-1\}\cup S_0)}{f_{\mathrm{M}}(j-1|\{k,j-1\}\cup S_0)}\geq\frac{f_{\mathrm{M}}(k|\{k,j-2\}\cup S_0)}{f_{\mathrm{M}}(j-2|\{k,j-2\}\cup S_0)}\geq\cdots\geq\frac{f_{\mathrm{M}}(k|\{k,k+1\}\cup S_0)}{f_{\mathrm{M}}(k+1|\{k,k+1\}\cup S_0)}=e^{\theta}.\end{aligned}$$

Note that the chain of inequalities above are all valid because j is the immediate successor of k in the display set S, i.e., $\{k+1,\ldots,j-1\}\cap S=\emptyset$.

To prove (24), define the set $C=\{\sigma:\sigma(k)<\sigma(j)\ \&\ \sigma(k)<\sigma(i),\forall i\in S_0\}=\{\sigma:\sigma(k)<\sigma(j)\}\cap B_k$ and $\tilde{C}=\{\sigma:\sigma(k)<\sigma(j-1)\ \&\ \sigma(k)<\sigma(i),\forall i\in S_0\}=\{\sigma:\sigma(k)<\sigma(j-1)\}\cap B_k$. Hence

$$\begin{aligned}f_{\mathrm{M}}(k|S)=&\sum_{\sigma\in C}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}\\ =&\sum_{\sigma\in C\cap\tilde{C}}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}+\sum_{\sigma\in C\setminus\tilde{C}}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}\\ =&\sum_{\sigma\in C\cap\tilde{C}}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}+\sum_{\sigma\in B_k:\sigma(j-1)<\sigma(k)<\sigma(j)}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}\\ \overset{(a)}{=}&\sum_{\sigma\in C\cap\tilde{C}}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}+\sum_{\sigma\in B_k:\sigma(j)<\sigma(k)<\sigma(j-1)}\frac{e^{-\theta d(\phi(\sigma),\sigma_0)}}{\Psi(\theta)}\\ =&\sum_{\sigma\in C\cap\tilde{C}}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}+e^{\theta}\sum_{\sigma\in B_k:\sigma(j)<\sigma(k)<\sigma(j-1)}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}\\ \geq&\sum_{\sigma\in C\cap\tilde{C}}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}+\sum_{\sigma\in B_k:\sigma(j)<\sigma(k)<\sigma(j-1)}\frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\sigma \in C \cap \tilde{C}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} + \sum_{\sigma \in \tilde{C} \setminus C} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} \\
&= \sum_{\sigma \in \tilde{C}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} = f_{\mathrm{M}}(k|\tilde{S}).
\end{aligned}
$$

In part (a) of the derivation above, we use the fact that the bijection restricted to the set $\{\sigma \in B_k : \sigma(j) < \sigma(k) < \sigma(j-1)\}$ is a bijection from $\{\sigma \in B_k : \sigma(j) < \sigma(k) < \sigma(j-1)\}$ to $\{\sigma \in B_k : \sigma(j-1) < \sigma(k) < \sigma(j)\}$.

Similarly, to prove (25), let $D = \{\sigma : \sigma(j) < \sigma(k)\} \cap B_j$ and $\tilde{D} = \{\sigma : \sigma(j-1) < \sigma(k)\} \cap B_{j-1}$. We have

$$
\begin{aligned}
f_{\mathrm{M}}(j|S) &= \sum_{\sigma \in D} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} \\
&= \sum_{\sigma \in D \cap \tilde{D}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} + \sum_{\sigma \in D \setminus \tilde{D}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} \\
&\overset{(a)}{=} \sum_{\sigma \in D \cap \tilde{D}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} + \sum_{\sigma \in \tilde{D} \setminus D} \frac{e^{-\theta d(\phi(\sigma),\sigma_0)}}{\Psi(\theta)} \\
&\overset{(b)}{=} \sum_{\sigma \in D \cap \tilde{D}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} + e^{-\theta} \sum_{\sigma \in \tilde{D} \setminus D} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} \\
&\le \sum_{\sigma \in D \cap \tilde{D}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} + \sum_{\sigma \in \tilde{D} \setminus D} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} \\
&= \sum_{\sigma \in \tilde{D}} \frac{e^{-\theta d(\sigma,\sigma_0)}}{\Psi(\theta)} = f_{\mathrm{M}}(j-1|\tilde{S}).
\end{aligned}
$$

In parts (a) and (b) of the derivations above, we note that

$$
D \setminus \tilde{D} = \{\sigma : \sigma(j) < \sigma(i) \text{ for all } i \in \{k\} \cup S_0 \text{ but } \exists i' \in \{k\} \cup S_0 \text{ s.t. } \sigma(j-1) > \sigma(i')\}
$$

and

$$
\tilde{D} \setminus D = \{\sigma : \sigma(j-1) < \sigma(i) \text{ for all } i \in \{k\} \cup S_0 \text{ but } \exists i' \in \{k\} \cup S_0 \text{ s.t. } \sigma(j) > \sigma(i')\}.
$$

Hence $D \setminus \tilde{D} \subseteq A^c$ and $\tilde{D} \setminus S \subseteq A$. Also, ϕ restricted to $\tilde{D} \setminus D$ is a bijection from $\tilde{D} \setminus D$ to $D \setminus \tilde{D}$ so that increases the number of inconsistencies by one.

Appendix B: On the Lower Bound of the Sample Complexity of any δ-accurate Policy

In this appendix, we will prove Theorem 1 in a slightly more general hypothesis testing framework, motivated by that of Chernoff (1959). Besides proving this result, we will also discuss in Section B.4 some concrete examples of alternatives ranking and selection problems that can be cast within our proposed hypothesis framework. Finally, we provide an example in Section B.5 on how to apply our framework in the context of dueling bandits.

B.1. A General Hypothesis Testing Setting

Let Θ be an arbitrary parameter space of possible states of the world and let $\theta^* \in \Theta$ be the true (unknown) state. We let $\mathcal{H} = \{H_1, H_2, \ldots, H_I\} \in 2^\Theta$ denote a collection of subsets of Θ, which we interpret as alternative "hypotheses" regarding the value of θ^*. For every $\theta \in \Theta$, we let $H(\theta) = \{H_i \colon \theta \in H_i\}$ and $\overline{H}(\theta) = \mathcal{H} \setminus H(\theta)$ be the set of true and false hypotheses, respectively, when $\theta^* = \theta$. We only make one assumption regarding the relationship between parameter space Θ and the hypothesis space $\mathcal{H}$: for all $\theta \in \Theta$, there is at least one hypothesis that is true under θ.

Assumption 1. *We assume that $H(\theta) \neq \emptyset$ for all $\theta \in \Theta$.*

The setup of Θ and $\mathcal{H}$ means that our hypothesis testing framework corresponds to a multiple composite hypothesis testing problem with hypotheses that are not mutually exclusive. More specifically, let denote $\Theta_i := \{\theta : H(\theta) = H_i\}$ to be the set of parameters that are consistent with hypothesis H_i. We allow $|\mathcal{H}| = I > 2$ (i.e. multiple), $|\Theta_i| = \infty$ for every i (i.e., composite) and do not require $\{\Theta_i\}_i$ to be mutually exclusive.

The decision-maker has access to a collection $\mathcal{S}$ of experiments that she can use to learn about the value of θ^* and decide which hypotheses are true or false. An experiment $S \in \mathcal{S}$ is a random variable taking values in an outcome space $\mathcal{X}$. (This outcome space could depend on S but we consider it as independent to ease notation). We let $\{f_\theta(X|S) : X \in \mathcal{X}\}$ denote the probability distribution over outcomes of experiment $S \in \mathcal{S}$ when $\theta^* = \theta$, where $f_\theta(X|S)$ is the probability of observing outcome X when experiment S is used. We make the following assumptions on $f_\theta(\cdot|S)$:

Assumption 2. *For all $\theta \in \Theta$,*

(B-1) (Probability Mass Function) *For any $S \in \mathcal{S}$, $\sum_{X \in \mathcal{X}} f_\theta(X|S) = 1$;*

(B-2) (Non-degeneracy) *For all $S \in \mathcal{S}$ and $X \in \mathcal{X}$, $f_\theta(X|S) > 0$;*

(B-3) (Identifiability) *For every $\theta \neq \theta' \in \Theta$, there exists $S \in \mathcal{S}$ and $X \in \mathcal{X}$ such that $f_\theta(X|S) \neq f_{\theta'}(X|S)$.*

Note that this assumption is weaker than the p-separability requirement that we impose on the set of $\mathcal{M}_p$ preferences in Definition 1. The assumption about identifiability is also weaker than typically assumed in the literature. For example, Chernoff (1959) would require for all $S \in \mathcal{S}$ and $\theta \neq \theta' \in \Theta$, there exists $X \in \mathcal{X}$ such that $f_\theta(X|S) \neq f_{\theta'}(X|S)$.

An admissible learning dynamic policy has three parts:

1. An *experimentation rule*, i.e., a sequence of probability distributions $\{\lambda_t \in \Delta(\mathcal{S})\}_{t=1}^\infty$, with each $\lambda_t \in \mathcal{S}$ being adapted to the filtration $\mathcal{F}_t := \sigma(S_1, X_1, \ldots, S_{t-1}, X_{t-1})$.
2. A *stopping rule*, i.e., an $\mathcal{F}_t$ stopping time τ that determines when to stop experimenting.
3. A *final selection rule*, i.e., $d_\tau \in [I]$ that identifies which hypothesis to recommend.

We let $\pi = (\{\lambda_t\}_{t=1}^{\infty}, \tau, d_\tau)$ denote an admissible policy. A parameter $\theta \in \Theta$ and an admissible policy π induce a probability distribution $\mathbb{P}_\theta^\pi(\cdot)$ over the path space $\{S_1, X_1, \dots, S_t, X_t\}_t$. We also denote by $\mathbb{E}_\theta^\pi[\cdot]$ the expectation operator under $\mathbb{P}_\theta^\pi(\cdot)$. We say that an admissible policy π is *δ-accurate* if experimentation terminates almost surely and the probability of selecting a false hypothesis is less than δ; that is, $\mathbb{P}_\theta^\pi(\tau < \infty) = 1$ and $\mathbb{P}_\theta^\pi(d_\tau \notin H(\theta)) \le \delta$ for any $\theta \in \Theta$.

Note that the top-ranked selection problem in the main body of the paper is a special case of above with the following characteristics: (1) The space of parameter is the set of p-separable models, $\Theta = \mathcal{M}_p$; (2) Each hypothesis H_i corresponds to a case in which product i is the top-ranked product and $H(f) = \{\sigma_f^{-1}(1)\}$; (3) The set of experiments $\mathcal{S}$ is the set of display sets; (4) The outcome space $\mathcal{X}$ of an experiment coincides with the set of versions, i.e., $\mathcal{X} = [K]$. Also, we note that the identifiability requirement in Assumption 2 is weaker than the p-Separability requirement in Definition 1.

B.2. Re-Statement of Theorem 1

Let us restate Theorem 1 in the context of the hypothesis testing framework above. Given any experiment $S \in \mathcal{S}$ and probability distribution $\lambda \in \Delta(\mathcal{S})$, we define the Kullback-Leibler divergence between two model parameters θ and θ' as:

$$D_S(\theta||\theta') := \sum_{k \in \mathcal{X}} f_\theta(k|S) \log \frac{f_\theta(k|S)}{f_{\theta'}(k|S)} \quad \text{and} \quad D_\lambda(\theta||\theta') := \sum_{S \in \mathcal{S}} D_S(\theta||\theta') \cdot \lambda(S), \tag{26}$$

respectively. Given any $\theta \in \Theta$, we define $\mathcal{A}(\theta) := \{\theta' \in \Theta : H(\theta) \subseteq \overline{H}(\theta')\}$ to be the set of parameters that do not agree with θ on any hypothesis that θ treats as true. We also introduce the *information* measure $I_*(\theta)$ to be the optimal value of the following Max-Min problem:

$$I_*(\theta) = \sup_{\lambda \in \Delta(\mathcal{S})} \inf_{\theta' \in \mathcal{A}(\theta)} D_\lambda(\theta||\theta'). \tag{27}$$

Theorem: (Lower Bound on $\mathbb{E}_\theta^\pi[\tau]$) *Let $\delta \in (0,1)$. For every δ-accurate policy π and $\theta \in \Theta$ such that $I_*(\theta) > 0$, we have*

$$\mathbb{E}_\theta^\pi[\tau] \ge \frac{kl(\delta, 1-\delta)}{I_*(\theta)} \qquad \text{and} \qquad \liminf_{\delta \to 0} \frac{\mathbb{E}_\theta^\pi[\tau]}{\log\left(\frac{1}{\delta}\right)} \ge \frac{1}{I_*(\theta)},$$

where $kl(\delta, 1-\delta) = \delta \log\left(\frac{\delta}{1-\delta}\right) + (1-\delta)\log\left(\frac{1-\delta}{\delta}\right)$.

B.3. Proof of Theorem 1

We let $\pi = (\{\lambda_t\}_{t=1}^{\infty}, \tau, d_\tau)$ be an arbitrary δ-accurate policy and let $\mathcal{F}_t$ denote the filtration generated by the sampling history $\mathcal{H}_t := \{S_1, X_1, \dots, S_t, X_t\}$. Without loss of generality, we will assume that $\mathbb{E}_\theta[\tau] < \infty$, for otherwise the theorem is trivially true.

For any state $\theta \in \Theta$ and any alternative state $\theta' \in \mathcal{A}(\theta)$ we define the log-likelihood function

$$L_t^{\theta,\theta'} := L^{\theta,\theta'}(\mathcal{H}_t) = \sum_{\ell=1}^{t} \log\left(\tfrac{f_\theta(X_\ell|S_\ell)}{f_{\theta'}(X_\ell|S_\ell)}\right).$$

As in Wald's sequential probability ratio test (SPRT), we expect that a δ-accurate policy will stop sampling when $L_t^{\theta,\theta'}$ exceeds an upper threshold under θ. In what follows, we formalize this intuition and show that

$$\mathbb{E}_\theta\left[L_\tau^{\theta,\theta'}\right] \geq kl(\delta, 1-\delta), \quad \forall \theta \in \Theta \text{ and } \forall \theta' \in \mathcal{A}(\theta). \tag{28}$$

The proof of (28) is done in two steps:

Step 1: From Lemma 19 (equation (19)) in Kaufmann et al. (2016), it follows that for any event $\mathcal{E} \in \mathcal{F}_\tau$ we have that

$$\mathbb{E}_\theta[L_\tau^{\theta,\theta'}|\mathcal{E}] \geq \log\left(\frac{\mathbb{P}_\theta(\mathcal{E})}{\mathbb{P}_{\theta'}(\mathcal{E})}\right).$$

Hence, for any partition $\mathcal{P}$ of events of $\mathcal{F}_\tau$, the previous inequality implies that

$$\mathbb{E}_\theta[L_\tau^{\theta,\theta'}] \geq \sum_{\mathcal{E}\in\mathcal{P}} \mathbb{P}_\theta(\mathcal{E}) \log\left(\frac{\mathbb{P}_\theta(\mathcal{E})}{\mathbb{P}_{\theta'}(\mathcal{E})}\right). \tag{29}$$

Step 2: Let us first suppose that $\delta \in (0, \frac{1}{2}]$ and consider the partition $\mathcal{P} = \{\mathcal{E}', \Omega \setminus \mathcal{E}'\}$, where $\mathcal{E}' := \{d_\tau \in H(\theta')\}$ is the event that the recommended hypothesis d_τ is true under θ'. Since $\theta' \in \mathcal{A}(\theta)$, $H(\theta) \subset \overline{H}(\theta')$, and hence $H(\theta') \subset \overline{H}(\theta)$. As a further consequence, $\mathcal{E}' \subset \{d_\tau \in \overline{H}(\theta)\}$, the event that the recommended hypothesis d_τ is false under θ. It follows that $\mathbb{P}_\theta(\mathcal{E}') \leq \delta$ and $\mathbb{P}_{\theta'}(\mathcal{E}') \geq 1-\delta$ since π is δ-accurate. As a result,

$$\begin{aligned}
\sum_{\mathcal{E}\in\mathcal{P}} \mathbb{P}_\theta(\mathcal{E}) \log\left(\frac{\mathbb{P}_\theta(\mathcal{E})}{\mathbb{P}_{\theta'}(\mathcal{E})}\right) &= \mathbb{P}_\theta(\mathcal{E}') \log\left(\frac{\mathbb{P}_\theta(\mathcal{E}')}{\mathbb{P}_{\theta'}(\mathcal{E}')}\right) + \left(1-\mathbb{P}_\theta(\mathcal{E}')\right) \log\left(\frac{1-\mathbb{P}_\theta(\mathcal{E}')}{1-\mathbb{P}_{\theta'}(\mathcal{E}')}\right) \\
&= \mathbb{P}_\theta(\mathcal{E}') \log\left(\mathbb{P}_\theta(\mathcal{E}')\right) + \left(1-\mathbb{P}_\theta(\mathcal{E}')\right) \log\left(1-\mathbb{P}_\theta(\mathcal{E}')\right) \\
&\quad - \Big[\mathbb{P}_\theta(\mathcal{E}') \log\left(\mathbb{P}_{\theta'}(\mathcal{E}')\right) + \left(1-\mathbb{P}_\theta(\mathcal{E}')\right) \log\left(1-\mathbb{P}_{\theta'}(\mathcal{E}')\right)\Big] \\
&\geq \delta \log\left(\tfrac{\delta}{1-\delta}\right) + (1-\delta) \log\left(\tfrac{1-\delta}{\delta}\right) = kl(\delta, 1-\delta).
\end{aligned} \tag{30}$$

The inequality is due to the conditions $\mathbb{P}_\theta(\mathcal{E}') \leq \delta$ and $\mathbb{P}_{\theta'}(\mathcal{E}') \geq 1-\delta$ and the fact that the negative entropy function $\delta \mapsto \left[\delta \log(\delta) + (1-\delta)\log(1-\delta)\right]$ decreases in the interval $(0, 1/2]$ and increases in the interval $[1/2, 1)$. We have thus verified our claim for $\delta \in (0, 1/2]$. The same argument can be made when $\delta \in [\frac{1}{2}, 1)$ by defining $\mathcal{E}' := \{d_\tau = \theta^{-1}(1)\}$.

Combining (29) and (30) we get (28).

Let us now express $\mathbb{E}_\theta\left[L_\tau^{\theta,\theta'}\right]$ in terms of the value of the Kullback-Leibler divergence $D_S(\theta||\theta')$. To this end, let $N_t(S) := \sum_{i=1}^{t} \mathbb{I}\{S_i = S\}$ be the number of times experiment S is used between time period 1 and t, under policy π. Also, for each $S \in \mathcal{S}$, we define a sequence $\{Y_\ell^S\}_{\ell \geq 1}$ of iid random variables with probability distribution $f_\theta(\cdot|S)$. It follows that

$$\begin{aligned}\mathbb{E}_\theta\left[L_\tau^{\theta,\theta'}\right] &= \mathbb{E}_\theta\left[\sum_{t=1}^{\tau} \log\left(\frac{f_\theta(X_t|S_t)}{f_{\theta'}(X_t|S_t)}\right)\right] = \mathbb{E}_\theta\left[\sum_{S\in\mathcal{S}}\sum_{t=1}^{\tau} \mathbb{I}\{S_t = S\}\log\left(\frac{f_\theta(X_t|S)}{f_{\theta'}(X_t|S)}\right)\right] \\ &= \sum_{S\in\mathcal{S}} \mathbb{E}_\theta\left[\sum_{\ell=1}^{N_\tau(S)} \log\left(\frac{f_\theta(Y_\ell^S|S)}{f_{\theta'}(Y_\ell^S|S)}\right)\right] = \sum_{S\in\mathcal{S}} \mathbb{E}_\theta[N_\tau(S)]\,\mathbb{E}_\theta\left[\log\left(\frac{f_\theta(Y_1^S|S)}{f_{\theta'}(Y_1^S|S)}\right)\right] \\ &= \sum_{S\in\mathcal{S}} \mathbb{E}_\theta[N_\tau(S)]\, D_S\left(\theta||\theta'\right), \end{aligned} \tag{31}$$

where the second to last equality follows from Wald's identity (recall that we have assumed that $\mathbb{E}_\theta[\tau] < \infty$ which implies that $\mathbb{E}_\theta[N_\tau(S)] < \infty$). Combining (28) and (31), we get that

$$kl(\delta, 1-\delta) \leq \inf_{\theta'\in\mathcal{A}(\theta)} \sum_{S\in\mathcal{S}} \mathbb{E}_\theta[N_\tau(S)]\, D_S\left(\theta||\theta'\right) \tag{32}$$

The final step of the proof of Theorem 1 is to show that the right-hand side of (32) is bounded above by $\mathbb{E}_\theta[\tau] I_*(\theta)$. Indeed,

$$\begin{aligned}\inf_{\theta'\in\mathcal{A}(\theta)} \sum_{S\in\mathcal{S}} \mathbb{E}_\theta[N_\tau(S)]\, D_S\left(\theta||\theta'\right) &= \mathbb{E}_\theta[\tau] \inf_{\theta'\in\mathcal{A}(\theta)} \sum_{S\in\mathcal{S}} \frac{\mathbb{E}_\theta[N_\tau(S)]}{\mathbb{E}_\theta[\tau]} D_S\left(\theta||\theta'\right) \\ &\leq \mathbb{E}_\theta[\tau] \sup_{\lambda\in\Delta(\mathcal{S})} \inf_{\theta'\in\mathcal{A}(\theta)} \sum_{S\in\mathcal{S}} \lambda(S)\, D_S\left(\theta||\theta'\right) = \mathbb{E}_\theta[\tau] I_*(\theta). \end{aligned} \tag{33}$$

The inequality in (33) follows from the fact that $\lambda(S) = \frac{\mathbb{E}_\theta[N_\tau(S)]}{\mathbb{E}_\theta[\tau]}$ is a feasible choice of $\Delta(\mathcal{S})$. Combining (32) and (33) we get that

$$\mathbb{E}_\theta[\tau] \geq \frac{kl(\delta, 1-\delta)}{I_*(\theta)},$$

which proves the first part of Theorem 1. The second part follows from noticing that $\liminf_{\delta\to 0} kl(\delta, 1-\delta)/\log(1/\delta) = 1$. ■

B.4. Application to Other Ranking-and-Selection Problems

The hypothesis testing framework presented in Section B.1 can accommodate a variety of problem formulations. For example, one could change the definition of $\mathcal{H}$ to capture different objectives. For example, in a *full ranking identification* problem, the company is interested in recovering the full ranking σ_* associated with the ground-truth preference f_*. To incorporate that, we can define the set of hypotheses $\{H_\sigma : \sigma \in \Sigma\}$, where H_σ denotes the subset of preferences f for which $\sigma_f = \sigma$. In a *"strong" top-k identification* problem, the company wishes to identify a collection of items S with size $k = |S|$ such that all the items in S are ranked higher than those in $[K] \setminus S$. To model this setting we define hypotheses $\{H_S : S \subseteq \mathcal{S},\ |S| = k\}$, where H_S is the subset of preferences f

such that $S = \{\sigma_f^{-1}(1), \ldots, \sigma_f^{-1}(k)\}$. In a *"weak" top-k identification* problem, the company wishes to identify a single item that is ranked k^{th} or higher. To model this problem, we use the set of hypotheses $\{H_i : i \in [K]\}$, where H_i is the subset of preferences f for which product i is ranked at least top k that is, $i \in \{\sigma_f^{-1}(1), \ldots, \sigma_f^{-1}(k)\}$.

Besides considering different ranking and selection objectives, one can also adapt our framework to (i) consider additional constraints on the set $\mathcal{S}$ of available display sets (e.g., let $\mathcal{S} = \{S : |S| \leq m\}$ to incorporate capacity constraints) or (ii) use Θ to capture different forms of information structure (e.g., let Θ to be a larger set than $\mathcal{M}_p$ to relax the p-separability constraint) or (iii) use $\mathcal{X}$ to capture different feedback structures (e.g., consumers provide full rankings rather than single choices).

While we have a relatively good understanding of the lower bound of sample complexity, we leave it an open problem as of whether an MTP-based sampling algorithm still achieves a sample complexity that matches the lower bound (asymptotically). Also, we are interested in understanding the structural properties of MTP under these different formulations. For instance, we conjecture that the randomization distribution, as a result of the new Max-Min problem, is sparse in general.

B.5. A Dueling Bandit Example

We conclude this appendix devoted to the lower bound in Theorem 1 with an example of how to apply the result in the context of a different ranking-and-selection problem. Specifically, we consider the problem of identifying the product with the highest Borda score studied by Jamieson et al. (2015) (see also Heckel et al. 2019). In this problem, only pairwise comparisons (i.e., dueling bandits) are considered and consumers' preferences are then represented by a matrix $P = [p_{ij}]$, where p_{ij} is the probability that version $i \in [K]$ is preferred over version $j \in [K]$ when both are displayed together. The Borda score of product i is defined by

$$s_i := \tfrac{1}{K-1} \sum_{j \neq i} p_{ij},$$

that is, s_i is the probability that version i is selected when displayed with another uniformly randomly selected version. Jamieson et al. (2015) considered the fixed-confidence identification problem of finding the version $i^* = \arg\max\{s_i : i \in [K]\}$ and derived the following lower bound on the sample complexity of any δ-accurate policy.

Theorem: (Jamieson et al. 2015) *Consider a comparison matrix P such that (i) item 1 is the Borda winner; (ii) $\frac{3}{8} \leq p_{ij} \leq \frac{5}{8}, \forall i, j \in [K]$, and (iii) $K \geq 3$. Then for $\delta \leq 0.15$, any δ-PAC dueling bandits algorithm π to find the Borda winner has*

$$\mathbb{E}_P^{\pi}[\tau] \geq \frac{1}{40} \left(\frac{K-2}{K-1} \right)^2 \left(\sum_{i=2}^{K} \frac{1}{(s_1 - s_i)^2} \right) \log \frac{1}{2\delta}.$$

The Borda winner identification problem in Jamieson et al. (2015) is closely related to our framework in Section B.1 with the following characteristics:

- The space of parameters is the set of comparison matrices $P = [p_{ij}]$, where the Borda winder is uniquely (and strictly) defined. That is, $\Theta^P = \{P \in \mathbb{R}^{K\times K}_{++} : p_{ij} + p_{ji} = 1$ and $\exists\ i_* \in [K]$ such that $s_{i_*} > \max_{j \neq i_*} s_j\}$. Let us denote $i_*(P)$ to be the Borda winner matrix P;
- Each hypothesis H_i corresponds to case in which item i is the best item. That is, $\mathcal{H} = [K]$ and $H(P) = \{i_*(P)\}$;
- The set of experiments $\mathcal{S}^P$ is the set of all (unordered) pairwise display sets. That is, $\mathcal{S}^P = \{\{i,j\} : i \neq j$ and $i,j \in [K]\}$;
- The stochastic comparison model is such that $f_P(i|\{i,j\}) = p_{ij}$ and $f_P(j|\{i,j\}) = p_{ji}$.

The specifications above exactly match the problem setup in Jamieson et al. (2015) except that we further allow the dueling bandits algorithm π to know the additional information that P is non-degenerate, i.e., $p_{ij} \neq \{0,1\}$ for all i,j. Within this setting, our lower bound stated in Section B.2 can be transferred to the following result:

Theorem: (Our lower bound in the setting of Jamieson et al. 2015) *Let $\delta \in (0,1)$ and $P \in \Theta^P$. Any δ-PAC dueling bandits algorithm π to find the Borda winner has*

$$\mathbb{E}^\pi_P[\tau] \geq \frac{kl(\delta, 1-\delta)}{\sup\limits_{\lambda \in \Delta(\mathcal{S}^P)} \quad \inf\limits_{\tilde{P}: i_*(\tilde{P}) \neq i_*(P)} \quad \sum_i \sum_{j>i} \lambda(\{i,j\})\ kl(p_{ij}, \tilde{p}_{ij})},$$

where $kl(\delta, 1-\delta) = \delta \log\left(\frac{\delta}{1-\delta}\right) + (1-\delta)\log\left(\frac{1-\delta}{\delta}\right)$.

We claim that our lower bound dominates the lower bound in Jamieson et al. (2015). On one hand, our lower bound applies to a larger collection of comparison matrices P and values of δ. On the other hand, our lower bound is tighter (i.e., larger) than that in Jamieson et al. (2015) whenever a direct comparison is possible. We formalize the second part of our claim in the result below.

Proposition 7. *Let $P \in \Theta^P$ be such that (i) item 1 is the Borda winner; (ii) $\frac{3}{8} \leq p_{ij} \leq \frac{5}{8}$, $\forall i,j \in [K]$, and (iii) $K \geq 3$. Then for $\delta \leq 0.15$,*

$$\frac{kl(\delta, 1-\delta)}{\sup\limits_{\lambda \in \Delta(\mathcal{S}^P)} \quad \inf\limits_{\tilde{P}: i_*(\tilde{P}) \neq i_*(P)} \quad \sum_i \sum_{j>i} \lambda(\{i,j\})\ kl(p_{ij}, \tilde{p}_{ij})} \geq \frac{1}{40}\left(\frac{K-2}{K-1}\right)^2 \left(\sum_{i\neq 1} \frac{1}{(s_1 - s_i)^2}\right) \log\frac{1}{2\delta}.$$

Proof of Proposition 7: Let$P \in \Theta^P$ be an arbitrary comparison matrix that satisfies the condition (i), (ii) and (iii) in the proposition. Also, select $\delta \leq 0.15$. Notice that $kl(\delta, 1-\delta) \geq \log\frac{1}{2\delta}$ when $\delta \leq 0.15$. Hence it suffices to show that

$$\sup_{\lambda \in \Delta(\mathcal{S}^P)} \quad \inf_{\tilde{P}: i_*(\tilde{P}) \neq i_*(P)} \quad \sum_i \sum_{j>i} \lambda(\{i,j\})\ kl(p_{ij}, \tilde{p}_{ij}) \leq \frac{1}{\frac{1}{40}\left(\frac{K-2}{K-1}\right)^2 \left(\sum_{i\neq 1}\frac{1}{(s_1 - s_i)^2}\right)}.$$

In order to show the inequality above, let us follow some the constructions in Jamieson et al. (2015). For every $b \in \{2,\dots,K\}$, we select the alternative comparison matrix

$$\tilde{p}^b_{ij} = \begin{cases} p_{ij} + \frac{K-1}{K-2}(s_1 - s_b) + \varepsilon & \text{if } i = b \text{ and } j \neq \{1,b\} \\ p_{ij} - \frac{K-1}{K-2}(s_1 - s_b) - \varepsilon & \text{if } j = b \text{ and } i \neq \{1,b\} \\ p_{ij} & \text{otherwise.} \end{cases}$$

In other words, $\tilde{P}^b$ is constructed so that b is the Borda winner under $\tilde{P}^b$ and $\tilde{P}^b$ differs from P only in the indices $\{(b,j) : j \neq \{1,b\}\}$. Moreover, let us pick a sufficiently small ε, so that $\tilde{P}^b \in \Theta^P$ and

$$0 < \max_{j \neq \{1,b\}} kl(p_{bj}, \tilde{p}'_{bj}) < 20 \left(\frac{K-1}{K-2}(s_1 - s_b) \right)^2 =: \mathfrak{C}_b.$$

We refer the reader to Equations (4)-(5) in Jamieson et al. (2015) for discussions on why we are able to select such an ε. With the construction above, we notice that

$$\begin{aligned}
& \sup_{\lambda \in \Delta(\mathcal{S}^P)} \inf_{\tilde{P}: i_*(\tilde{P}) \neq i_*(P)} \sum_i \sum_{j>i} \lambda(\{i,j\})\ kl(p_{ij}, \tilde{p}_{ij}) \\
= & \sup_{\lambda \in \Delta(\mathcal{S}^P)} \min_{b \in \{2,\dots,K\}} \inf_{\tilde{P}: i_*(\tilde{P}) = b} \sum_i \sum_{j>i} \lambda(\{i,j\})\ kl(p_{ij}, \tilde{p}_{ij}) \\
\leq & \sup_{\lambda \in \Delta(\mathcal{S}^P)} \min_{b \in \{2,\dots,K\}} \sum_i \sum_{j>i} \lambda(\{i,j\})\ kl(p_{ij}, \tilde{p}^b_{ij}) \\
= & \sup_{\lambda \in \Delta(\mathcal{S}^P)} \min_{b \in \{2,\dots,K\}} \sum_{j \neq \{1,b\}} \lambda(\{b,j\})\ kl(p_{bj}, \tilde{p}^b_{bj}) \\
\leq & \sup_{\lambda \in \Delta(\mathcal{S}^P)} \min_{b \in \{2,\dots,K\}} \sum_{j \neq \{1,b\}} \lambda(\{b,j\})\ \mathfrak{C}_b \\
\leq & \sup_{x_b \geq 0:\ \sum_{b=2}^K x_b \leq 2} \min_{b \in \{2,\dots,K\}} x_b\ \mathfrak{C}_b \qquad [x_b := \sum_{j \neq \{1,b\}} \lambda(\{b,j\})] \\
= & \frac{2}{\frac{1}{\mathfrak{C}}_2 + \cdots \frac{1}{\mathfrak{C}_K}} = \frac{1}{\frac{1}{40} \left(\frac{K-2}{K-1} \right)^2 \left(\sum_{i=2}^K \frac{1}{(s_1 - s_i)^2} \right)}.
\end{aligned}$$

This finishes the proof. ■

Appendix C: Proof of Theorem 2

This proof draws inspiration from Chernoff (1959) and is carefully adapted to the current setting.

C.1. Preliminaries

Let us introduce some notations. Define $\overline{\mathcal{M}}_p^{\mathrm{F}}(f) := \overline{\mathcal{M}}_p(f) \cap \mathcal{M}_p^{\mathrm{F}}$. Also, let us introduce $I_*^F(f) := \max_{\lambda\in\Delta(\mathcal{S})} \min_{\bar{f}\in\overline{\mathcal{M}}_p^{\mathrm{F}}(f)} D_\lambda\left(f||\bar{f}\right) \geq \max_{\lambda\in\Delta(\mathcal{S})} \min_{\bar{f}\in\overline{\mathcal{M}}_p(f)} D_\lambda\left(f||\bar{f}\right) = I_*(f)$. Let $\lambda_*^F(f) \in \arg\max_{\lambda\in\Delta(\mathcal{S})} \min_{\bar{f}\in\overline{\mathcal{M}}_p^{\mathrm{F}}(f_t^F)} D_\lambda\left(f_t^F||\bar{f}\right)$. If there are multiple optimal solutions, we pick a arbitrary but fixed rule to break ties (e.g., in the lexicographical order). The detailed description of the policy $\hat{\pi}$ is summarized in Algorithm 2.

Algorithm 2 Policy $\hat{\pi}$

INPUT: $\mathcal{M}_p^{\mathrm{F}}$.

STEP 1: At each epoch t, given the history of votes $(S_1, X_1, \ldots, S_t, X_t)$, compute the most likely consensus preference by solving the MLE problem under $\mathcal{M}_p^{\mathrm{F}}$

$$f_t^F \in \arg\max_{f\in\mathcal{M}_p^{\mathrm{F}}} \sum_{\ell=1}^{t} \log f(X_\ell|S_\ell). \tag{34}$$

We break ties arbitrarily if the arg max in (MLE) is not a singleton.

STEP 2: Update the value of the generalized log-likelihood ratio process

$$\mathcal{L}_t^F = \min_{\bar{f}\in\overline{\mathcal{M}}_p^{\mathrm{F}}(f_t^F)} L_t^{f_t^F,\bar{f}}. \tag{35}$$

If $\mathcal{L}_t^F \geq \beta^F := \log(|\mathcal{M}_p^{\mathrm{F}}|) + \log\left(\frac{1}{\delta}\right)$ then stop and select the top-ranked version according to f_t^F that is, $\sigma_{f_t^F}^{-1}(1)$. Otherwise, go to Step 3.

STEP 3: If t is a perfect square number (i.e., there exists an integer i such that $t = i^2$), then pick λ_t to be the uniform distribution over all display sets, i.e., $\lambda_t(S) = \frac{1}{|\mathcal{S}|}$ for all $S \in \mathcal{S}$. Otherwise, pick $\lambda_t = \lambda_*^F(f_t^F)$. Randomly select a set using the probability distribution λ_t to be displayed to the next consumer and record her choice X_{t+1}. Go to Step 1 and iterate. □

C.2. Main Body of Proof

Proof of Theorem 2. Let us break the proof into two steps.

Step 1. We claim that for all $f \in \mathcal{M}_p^F$, (i) $\mathbb{E}_f[\tau] < +\infty$ for all $\delta \in (0,1)$; and (ii) $\limsup_{\delta\downarrow 0} \frac{\mathbb{E}_f^{\hat{\pi}}[\tau]}{\log\left(\frac{1}{\delta}\right)} \leq \frac{1}{I_*^F(f)}$, which implies that $\limsup_{\delta\downarrow 0} \frac{\mathbb{E}_f^{\hat{\pi}}[\tau]}{\log\left(\frac{1}{\delta}\right)} \leq \frac{1}{I_*(f)}$.

We invoke the following auxiliary lemma that gives an upper bound to the tail probability of τ, i.e., $\mathbb{P}_f(\tau \geq t)$. Section C.3 contains proof of this auxiliary lemma.

Lemma 1. *For all $0<\epsilon<1$, there exists a convergent series $\{\rho_t\}_t>0$ (i.e. $\sum_{t=1}^{\infty}\rho_t<\infty$), which is independent of δ, such that for every $\delta\in(0,1)$, $f\in\mathcal{M}_p^{\mathrm{F}}$ and $t\geq M(\delta):=\frac{1+\epsilon}{I_*^F(f)}\log\frac{1}{\delta}$,*

$$\mathbb{P}_f(\tau\geq t)\leq\rho_t. \tag{36}$$

Lemma 1 above is sufficient for Step 1. To see why, pick an arbitrary $\epsilon\in(0,1)$, $f\in\mathcal{M}_p^F$, and $\{\rho_t\}_t$ as stated in the lemma.

$$\begin{aligned}
\mathbb{E}_f[\tau]=\sum_{t=1}^{\infty}\mathbb{P}_f(\tau\geq t)\leq\sum_{t=1}^{M(\delta)}\mathbb{P}_f(\tau\geq t)+\sum_{t=M(\delta)}^{\infty}\mathbb{P}_f(\tau\geq t) &\qquad [M(\delta)\text{ is defined in Lemma 1}]\\
\leq\sum_{t=1}^{M(\delta)}1+\sum_{t=M(\delta)}^{\infty}\mathbb{P}_f(\tau\geq t) &\qquad [\mathbb{P}_f(\tau\geq t)\leq 1]\\
\leq M(\delta)+\sum_{t=M(\delta)}^{\infty}\rho_t &\qquad [\text{Lemma 1}]\\
\leq\tfrac{1+\epsilon}{I_*^F(f)}\log\tfrac{1}{\delta}+C', &\qquad [C':=\sum_{t=1}^{\infty}\rho_t<\infty]
\end{aligned}$$

Due to Lemma 1, C' is a finite constant independent of δ. Hence $\frac{\mathbb{E}_f[\tau]}{\log\frac{1}{\delta}}\leq\frac{1+\epsilon}{I_*^F(f)}+\frac{C'}{-\log\delta}<+\infty$. Moreover, $\frac{\mathbb{E}_f[\tau]}{\log\frac{1}{\delta}}\leq\frac{1+2\epsilon}{I_*^F(f)}$ for sufficiently small δ. Take $\epsilon,\delta\to 0$, and we finish the proof.

<u>Step 2.</u> We claim that $\hat{\pi}$ is $\delta(\mathcal{M}_p^{\mathrm{F}})$-accurate for every $\delta\in(0,1)$.

Given any $f\in\mathcal{M}_p^F$, $\mathbb{P}_f(\tau<\infty)=1$ due to Lemma 1. Given every $\bar{f}\in\overline{\mathcal{M}}_p^F(f)$, let $\mathcal{E}_{\bar{f}}$ be the event that the policy $\hat{\pi}$ terminates with estimated state $f_\tau^F=\bar{f}$ (which produces a mistake). Notice that

$$\begin{aligned}
\mathbb{P}_f(\mathcal{E}_{\bar{f}})=&\mathbb{E}_{\bar{f}}[\mathbb{I}\{\mathcal{E}_{\bar{f}}\}\exp(-L_\tau^{\bar{f},f})] &\qquad [\text{change-of-measure}]\\
\leq&\tfrac{\delta}{|\mathcal{M}_p^{\mathrm{F}}|}. &\qquad [\text{Due to the stopping rule }\tau,\ L_\tau^{\bar{f},f}\geq\beta^F=\log(|\mathcal{M}_p^{\mathrm{F}}|)+\log\left(\tfrac{1}{\delta}\right)]
\end{aligned}$$

As a result, $\mathbb{P}_f(d_\tau\neq\sigma_f^{-1}(1))=\sum_{\bar{f}\in\overline{\mathcal{M}}_p^F(f)}\mathbb{P}_f(\mathcal{E}_{\bar{f}})\leq\frac{|\overline{\mathcal{M}}_p^F(f)|}{|\mathcal{M}_p^{\mathrm{F}}|}\delta\leq\delta$. ■

Remark 4. The particular value of β^F is only used in Step 2 above. As a result, we can change β^F to $\tilde{C}^F+\log(1/\delta)$ for an arbitrary finite constant $\tilde{C}^F$ independent of δ, without affecting the conclusion in Step 1 above, i.e., $\limsup_{\delta\downarrow 0}\frac{\mathbb{E}_f^{\hat{\pi}}[\tau]}{\log\left(\frac{1}{\delta}\right)}\leq\frac{1}{I_*^F(f)}$ (also see Step 3 of the proof of Lemma 3).

C.3. Proof of the Auxiliary Lemma 1

Before we prove Lemma 1, we first prove two technical lemmas below.

Lemma 2. *Given any $\mathbb{Z}_+$ valued random variable X and probability measure μ, the following are equivalent:*

1. $\mathbb{E}_\mu[X^2]<\infty$;
2. $\sum_{s=1}^{\infty}s\mathbb{P}_\mu(X\geq s)<\infty$;

3. $\sum_{t=1}^{\infty}\sum_{s=t}^{\infty}\mathbb{P}_\mu(X\geq s)<\infty$.

Proof. We prove this lemma by applying Fubini's Theorem twice. Note that all the items involved in the summation are nonnegative,

$$\sum_{t=1}^{\infty}\sum_{s=t}^{\infty}\mathbb{P}_\mu(X\geq s)=\sum_{s=1}^{\infty}\sum_{t=1}^{s}\mathbb{P}_\mu(X\geq s)=\sum_{s=1}^{\infty}s\mathbb{P}_\mu(X\geq s)=\sum_{s=1}^{\infty}s\sum_{z=s}^{\infty}\mathbb{P}_\mu(X=z)=\sum_{z=1}^{\infty}\left(\sum_{s=1}^{z}s\right)\mathbb{P}_\mu(X=z)$$
$$=\sum_{z=1}^{\infty}\tfrac{z(z+1)}{2}\mathbb{P}_\mu(X=z)=\tfrac{\mathbb{E}_\mu[X^2]}{2}+\tfrac{\mathbb{E}_\mu[X]}{2},$$

and hence the statement of the lemma follows. ■

Let $\hat{\tau}:=\max\{t:f_t^F\neq f\}$, the last time f_t^F is not equal to f. Lemma 3 demonstrates the speed of convergence of the estimated preference f_t^F to f.

LEMMA 3. *For all $f\in\mathcal{M}_p^F$, there exists $C,\epsilon>0$, independent of δ, such that for every $t\in\{1,2,\dots\}$, $\mathbb{P}_f(f_t^F\neq f)\leq Ce^{-\epsilon\sqrt{t}}$. As a result, $\mathbb{E}_f[\hat{\tau}^2]<+\infty$.*

Proof of Lemma 3. Pick an arbitrary $\bar{f}\in\mathcal{M}_p^F\setminus\{f\}$. For all $t\in\mathbb{Z}_+$, $D_{\lambda_t}\left(f||\bar{f}\right)\geq 0$. Moreover, since $f\neq\bar{f}$, there exists $S\in\mathcal{S}$ such that $D_S\left(f||\bar{f}\right)>0$, and thus $D_{\lambda_t}\left(f||\bar{f}\right)>0$ if t is a perfect square number, due to the construction of Step 3 in the policy $\hat{\pi}$. We refer to the same argument in Lemma 1 of Chernoff (1959), and conclude that there exists $C_{\bar{f}},\epsilon_{\bar{f}}>0$ such that $\mathbb{P}_f(f_t^F=\bar{f})\leq C_{\bar{f}}e^{-\epsilon_{\bar{f}}\sqrt{t}}$, for every $t\in\{1,2,\dots\}$. It now suffices to pick $\epsilon:=\min\{\epsilon_{\bar{f}}:\bar{f}\in\overline{\mathcal{M}}_p^F\}$ and $C:=|\mathcal{M}_p^F|\max\{C_{\bar{f}}:\bar{f}\in\overline{\mathcal{M}}_p^F\}$ to satisfy our claim. Noting that $e^{-\epsilon\sqrt{t}}$ decays faster than $1/t^\alpha$ for all $\alpha>0$,

$$\sum_{t=1}^{\infty}t\mathbb{P}_f(\hat{\tau}\geq t)\leq\sum_{t=1}^{\infty}t\left(\sum_{\ell=t}^{\infty}\mathbb{P}_f(f_t^F\neq f)\right)\leq C\sum_{t=1}^{\infty}t\left(\sum_{\ell=t}^{\infty}e^{-\epsilon\sqrt{\ell}}\right)<+\infty.$$

Hence, $\mathbb{E}_f[\hat{\tau}^2]<+\infty$. In addition, the quantity does not depend on δ, because δ is only used for the stopping rule. ■

Proof of Lemma 1. Fix the $\epsilon\in(0,1)$ and $f\in\mathcal{M}_p^F$ stated in Lemma 1 throughout the proof. Also let $\beta^F:=\log(|\mathcal{M}_p^F|)+\log\left(\frac{1}{\delta}\right)$ for ease of notation. We split the discussion into three parts.

Part 1. We give an upper bound of the tail probability $\mathbb{P}_f(\tau\geq T)$. This upper bound implies that it suffices to show that the probability $\mathbb{P}_f\left(L_t^{f,\bar{f}}<\beta^F\right)$ is small (uniformly in $\bar{f}\in\overline{\mathcal{M}}_p^F(f)$) when $t\geq M(\delta)$.

For every $\bar{f}\in\overline{\mathcal{M}}_p^F(f)$, we introduce $\tau_{\bar{f}}:=\max\{t:L_t^{f,\bar{f}}<\beta^F\}\in\{0,1,\dots,\infty\}$ to be the final time that $L_t^{f,\bar{f}}$ is below β^F. According to the definition of τ under policy $\hat{\pi}$,

$$\mathbb{P}_f(\tau\geq t)=\mathbb{P}_f\left(\min\left\{\ell:\min_{\bar{f}\overline{\mathcal{M}}_p^F(f_\ell^F)}L_\ell^{f_\ell^F,\bar{f}}\geq\beta^F\right\}\geq t\right)$$

$$
\begin{aligned}
&\le \mathbb{P}_f\left(\max\left\{\ell : \min_{\bar f \in \overline{\mathcal{M}}_p^F(f_\ell^F)} L_\ell^{f_\ell^F,\bar f} < \beta^F\right\} \ge t-1\right)\\
&\le \sum_{\ell=t-1}^{\infty} \mathbb{P}_f\left(\min_{\bar f \in \overline{\mathcal{M}}_p^F(f_t^F)} L_\ell^{f_\ell^F,\bar f} < \beta^F\right)\\
&= \sum_{\ell=t-1}^{\infty} \mathbb{P}_f\left(\min_{\bar f \in \overline{\mathcal{M}}_p^F(f_\ell^F)} L_\ell^{f_\ell^F,\bar f} < \beta^F \text{ and } f_\ell^F = f\right) + \sum_{\ell=t-1}^{\infty} \mathbb{P}_f\left(\min_{\bar f \in \overline{\mathcal{M}}_p^F(f_\ell^F)} L_\ell^{f_\ell^F,\bar f} < \beta^F \text{ and } f_\ell^F \ne f\right)\\
&\le \sum_{\ell=t-1}^{\infty} \mathbb{P}_f\left(\min_{\bar f \in \overline{\mathcal{M}}_p^F(f)} L_\ell^{f,\bar f} < \beta^F\right) + \sum_{\ell=t-1}^{\infty} \mathbb{P}_f(f_\ell^F \ne f)\\
&\le \sum_{\ell=t-1}^{\infty} \mathbb{P}_f\left(L_\ell^{f,\bar f} < \beta^F \text{ for some } \bar f \in \overline{\mathcal{M}}_p^F(f)\right) + \sum_{\ell=t-1}^{\infty} \mathbb{P}_f(f_\ell^F \ne f)\\
&\le \sum_{\bar f \in \overline{\mathcal{M}}_p^F(f)} \sum_{\ell=t-1}^{\infty} \mathbb{P}_f\left(L_\ell^{f,\bar f} < \beta^F\right) + \sum_{\ell=t-1}^{\infty} \mathbb{P}_f(f_\ell^F \ne f).
\end{aligned}
$$

We claim that it suffices to construct a sequence $\{\tilde\rho_t\}_{t=0}^\infty$, independent of δ, such that the following two conditions hold:

$$
\begin{aligned}
&\sum_{t=1}^{\infty}\sum_{\ell=t-1}^{\infty} \tilde\rho_\ell < \infty,\\
&\mathbb{P}_f\left(L_t^{f,\bar f} < \beta^F\right) \le \tilde\rho_t, \text{ for every } \delta \in (0,1), t \ge M(\delta) \text{ and } \bar f \in \overline{\mathcal{M}}_p^F.
\end{aligned}
\tag{37}
$$

To see why, recall that our goal is to find $\{\rho_t\}_t$ such that (i) $\sum_t \rho_t < \infty$ and (ii) $\mathbb{P}_f(\tau \ge t) \le \rho_t$ for all $\delta \in (0,1)$ and $t \ge M(\delta)$. Given construction of $\{\tilde\rho_t\}_{t=0}^\infty$, it is easy to verify that $\rho_t := \sum_{\bar f \in \overline{\mathcal{M}}_p^F(f)} \sum_{\ell=t-1}^\infty \tilde\rho_t + \sum_{\ell=t-1}^\infty \mathbb{P}_f(f_\ell^F \ne f)$ satisfies our needs. In fact, the verification follows from the development above as well as the following two facts: (i) $\overline{\mathcal{M}}_p^F(f)$ is a finite set; and (ii) $\sum_{t=1}^\infty \sum_{\ell=t-1}^\infty \mathbb{P}_f(f_\ell^F \ne f) \le \sum_{t=1}^\infty \sum_{\ell=t-1}^\infty C e^{-\epsilon\sqrt{\ell}} < +\infty$ due to Lemma 3.

Part 2. We estimate the log-likelihood ratio process $L_t^{f,\bar f}$ to help us construct $\{\tilde\rho_t\}_{t=0}^\infty$.

Following a similar idea in Chernoff (1959) (Lemma 2 as well as Footnote 7), we may separate the likelihood ratio process into three parts: (i) the "noise" part from the choices (conditional on the sequence of display sets); (ii) the "noise" part from the randomness of displaying sets (since the algorithm decides which display set to offer at each epoch based on historical data and possible randomization); and (iii) the "deterministic" part (which captures the long-run average growth rate of the process). Formally, we introduce the one-shot log-likelihood ratio function $L^{f,\bar f}(X,S) : (X,S) \mapsto \log \frac{f(X|S)}{\bar f(X|S)}$. Also, let us write $\lambda_*^F = \lambda_*^F(f)$ for shorthand notation. Observe that

$$
\begin{aligned}
L_t^{f,\bar f} = \sum_{\ell=1}^t L^{f,\bar f}(X_\ell, S_\ell) = &\underbrace{\sum_{\ell=1}^t \left[L^{f,\bar f}(X_\ell,S_\ell) - D_{S_\ell}\left(f||\bar f\right)\right]}_{A} + \underbrace{\sum_{\ell=1}^t \left[D_{S_\ell}\left(f||\bar f\right) - D_{\lambda_*^F}\left(f||\bar f\right)\right]}_{B}\\
&+ \underbrace{t \cdot D_{\lambda_*^F}\left(f||\bar f\right)}_{C}
\end{aligned}
$$

Here, Part A corresponds to the noise part from the choices, Part B corresponds to the noise part from display sets, and Part C corresponds to the deterministic part. We will show that both Part A and B diverge sub-linearly in time (in the sense that the tail probabilities $\mathbb{P}_f(A \le -\varepsilon t)$ and $\mathbb{P}_f(B \le -\varepsilon t)$ decay fast in t for all $\varepsilon > 0$) while Part C grows at least as fast as the deterministic linear function $I_*^F(f)\,t$.

We claim that Part A is a $\mathbb{P}_f$-martingale with bounded differences, so that it diverges sublinearly. We first verify the martingale property. Observe that for every $t \ge 0$,

$$\begin{aligned}
&\mathbb{E}_f\left[L^{f,\bar{f}}(X_{t+1}, S_{t+1})|\mathcal{F}_t\right] \\
=&\mathbb{E}_f\left[\mathbb{E}_f\left[\log\left(\tfrac{f(X_{t+1}|S_{t+1})}{\bar{f}(X_{t+1}|S_{t+1})}\right)\middle|\mathcal{F}_t, S_{t+1}\right]\middle|\mathcal{F}_t\right] && \text{[tower property]} \\
=&\mathbb{E}_f\left[\mathbb{E}_f\left[\log\left(\tfrac{f(X_{t+1}|S_{t+1})}{\bar{f}(X_{t+1}|S_{t+1})}\right)\middle|S_{t+1}\right]\middle|\mathcal{F}_t\right] && \text{[independence of } X_{t+1} \text{ conditional on } S_{t+1}\text{]} \\
=&\mathbb{E}_f\left[D_{S_{t+1}}\left(f||\bar{f}\right)\middle|\mathcal{F}_t\right].
\end{aligned}$$

That implies that $\mathbb{E}_f\left[L^{f,\bar{f}}(X_{t+1}, S_{t+1}) - D_{S_{t+1}}\left(f||\bar{f}\right)\middle|\mathcal{F}_t\right] = 0$, for every $t \ge 0$. Hence Part A is a $\mathbb{P}_f$-martingale. It has bounded difference, because for every $S \in \mathcal{S}$ and $k \in S$,

$$\log \tfrac{f(k|S)}{\bar{f}(k|S)} \le \overline{\mathcal{L}} := \max_{S' \in \mathcal{S}; k' \in S', f', \bar{f}' \in \mathcal{M}_p^F} \log \tfrac{f'(k'|S')}{\bar{f}'(k'|S')} \overset{(A-1)}{<} +\infty.$$

That means $D_S\left(f||\bar{f}\right)$ is uniformly bounded by $\overline{\mathcal{L}}$ as well. Part A diverges sublinearly in the sense that due to Azuma's inequality (Chung and Lu 2006), for every $\epsilon_2 > 0$

$$\mathbb{P}_f(A \le -\epsilon_2 t) \le \exp\left(\tfrac{-\epsilon_2^2 t}{2\overline{\mathcal{L}}^2}\right). \tag{38}$$

To estimate Part B, recall $\hat{\tau} = \max\{t : f_t^F \ne f\}$, the last time the estimated ranking f_t^F differs from f. In other words, for all $t \ge \hat{\tau}+1$, $f_t^F = f$. For the sake of analysis, let $\{\tilde{S}_\ell\}_\ell$ be a sequence of i.i.d. $\mathcal{S}$-valued random variables with distribution λ_*^{OA} such that $\tilde{S}_\ell = S_\ell$ for all $t \ge \ell \ge \hat{\tau}+1$. We may write Part B as

$$\begin{aligned}
B = &\underbrace{\sum_{\ell=1}^{t}\left[D_{\tilde{S}_\ell}\left(f||\bar{f}\right) - D_{\lambda_*^F}\left(f||\bar{f}\right)\right]}_{B_1} + \underbrace{\sum_{\ell=1}^{\hat{\tau}\wedge t}\left[D_{S_\ell}\left(f||\bar{f}\right) - D_{\tilde{S}_\ell}\left(f||\bar{f}\right)\right]}_{B_2} + \\
&\underbrace{\sum_{\ell:\hat{\tau}+1 \le \ell \le t, \sqrt{\ell} \in \mathbb{Z}}\left[D_{S_\ell}\left(f||\bar{f}\right) - D_{\tilde{S}_\ell}\left(f||\bar{f}\right)\right]}_{B_3}
\end{aligned}$$

We will show that all of Parts B_1, B_2, B_3 diverge sublinearly: Part B_1 is a sum of IID random variables with mean zero; both Part B_2 and B_3 take negligible fractions for the time epochs. Specifically, B_1 is a partial sum of a sequence of i.i.d random variables $\left\{D_{\tilde{S}_\ell}\left(f||\bar{f}\right) - D_{\lambda_*^F}\left(f||\bar{f}\right)\right\}_\ell$, where $\tilde{S}_\ell$ is a $\mathcal{S}$-valued random variable with distribution λ_*^F. Hence B_1 is a martingale with (uniformly) bounded differences $2\overline{\mathcal{L}}$. Due to Azuma's inequality, for every $\epsilon_2 > 0$,

$$\mathbb{P}_f(B_1 \le -\epsilon_2 t) \le \exp\left(\frac{-\epsilon_2^2 t}{2\overline{\mathcal{L}}^2}\right). \tag{39}$$

Moreover, again invoking the uniform boundedness of $D_S\left(f||\bar{f}\right)$, $B_2 \ge \sum_{\ell=1}^{\hat{\tau}\wedge t}(-2\overline{\mathcal{L}}) \ge -2\,\overline{\mathcal{L}}\,\hat{\tau}$. As a result,

$$\mathbb{P}_f(B_2 \le -\epsilon_2 t) \le \mathbb{P}_f(-2\,\overline{\mathcal{L}}\,\hat{\tau} \le -\epsilon_2 t) \le \mathbb{P}_f\left(\hat{\tau} \ge \frac{\epsilon_2 t}{2\overline{\mathcal{L}}}\right). \tag{40}$$

We may also obtain a bound for Part B_3, which also diverges sublinearly:

$$\mathbb{P}_f(B_3 \le -\epsilon_2 t) \le \mathbb{P}_f(-2\overline{\mathcal{L}}\sqrt{t} \le -\epsilon_2 t) = 0 \text{ for all } t \ge \sqrt{2\overline{\mathcal{L}}/\epsilon_2}. \tag{41}$$

To estimate Part C, simply observe that

$$C = D_{\lambda_*^F}\left(f||\bar{f}\right)t \ge \min_{f'\overline{\mathcal{M}}_p^F(f)} D_{\lambda_*^F}\left(f||f'\right)t = I_*^F(f)\,t. \tag{42}$$

Part 3. We use the estimates of $L_t^{f,\bar{f}}$, i.e. (38) through (42), to construct proper $\{\tilde{\rho}\}_{t=0}^{\infty}$ that satisfies (37).

We pick $\epsilon_2 := \frac{\epsilon I_*}{8(1+\epsilon)}$, $C_F := \log(|\mathcal{M}_p^F|)$, and define

$$\tilde{\rho}_t = \begin{cases} 1, & t \le \frac{2(1+\epsilon)C_F}{\epsilon I_*} \vee \sqrt{2\overline{\mathcal{L}}/\epsilon_2} \\ \mathbb{P}_f\left(\hat{\tau} \ge \frac{\epsilon_2 t}{2\overline{\mathcal{L}}}\right) + 2\exp\left(\frac{-\epsilon_2^2 t}{2\overline{\mathcal{L}}^2}\right), & t > \frac{2(1+\epsilon)C_F}{\epsilon I_*} \vee \sqrt{2\overline{\mathcal{L}}/\epsilon_2}. \end{cases} \tag{43}$$

Due to Lemma 3, the construction of $\tilde{\rho}_t$ is independent of δ. Let us verify the first line in (37):

$$\sum_{T=1}^{\infty}\sum_{t=T-1}^{\infty}\tilde{\rho}_t = \sum_{t=0}^{\infty}(t+1)\tilde{\rho}_t \le \sum_{t=1}^{\frac{2(1+\epsilon)C_F}{\epsilon I_*}\vee\sqrt{2\overline{\mathcal{L}}/\epsilon_2}}(t+1) + \sum_{t=1}^{\infty}(t+1)\mathbb{P}_f\left(\hat{\tau} \ge \frac{\epsilon_2 t}{2\overline{\mathcal{L}}}\right) + \sum_{t=1}^{\infty}(t+1)\cdot 2\exp\left(\frac{-\epsilon_2^2}{2\overline{\mathcal{L}}^2}\cdot t\right)$$

Of the three items on the right-hand side of the inequality above, the first term is a finite sum; the second term is finite because of the finite second moment of $\hat{\tau}$ in Lemma 3 as well as Lemma 2; and the third item is finite because $\exp\left(\frac{-\epsilon_2^2}{2\overline{\mathcal{L}}^2}\cdot t\right)$ decays exponentially fast in t.

Let us verify the second line in (37). Pick any $\delta \in (0,1), t \ge M(\delta)$ and $\bar{f} \in \overline{\mathcal{M}}_p^F$. Assume that $t \ge \frac{2(1+\epsilon)C_F}{\epsilon I_*^F(f)} \vee \sqrt{2\overline{\mathcal{L}}/\epsilon_2}$, on top of $t \ge M(\delta)$, without loss of generality in light of (43).

$$\begin{aligned} &\mathbb{P}_f\left(L_t^{f,\bar{f}} < \beta^F\right) = \mathbb{P}_f(A + B_1 + B_2 + B_3 + C < \beta^F) \\ \le&\mathbb{P}_f\left(A + B_1 + B_2 + B_3 + tI_*^F(f) < C_F + \log\tfrac{1}{\delta}\right) && \text{[due to (42)]} \\ \le&\mathbb{P}_f\left(A + B_1 + B_2 + B_3 + tI_*^F(f) < C_F + \tfrac{I_*^F(f)}{1+\epsilon}t\right) && [t \ge M(\delta) = \tfrac{1+\epsilon}{I_*^F(f)}\log\tfrac{1}{\delta}] \\ =&\mathbb{P}_f\left(A + B_1 + B_2 + B_3 < C_F - \tfrac{\epsilon I_*^F(f)}{1+\epsilon}t\right) \end{aligned}$$

$$
\begin{aligned}
&\le \mathbb{P}_f\left(A+B_1+B_2+B_3<-\tfrac{\epsilon I_*^F(f)}{2(1+\epsilon)}t\right) && \left[t\ge \tfrac{2(1+\epsilon)C_F}{\epsilon I_*^F(f)} \Rightarrow C_F<\tfrac{\epsilon I_*^F(f)}{2(1+\epsilon)}t\right]\\
&\le \mathbb{P}_f\left(A+B_1+B_2+B_3<-4\epsilon_2 t\right) && \left[\epsilon_2\le \tfrac{\epsilon I_*^F(f)}{8(1+\epsilon)}\right]\\
&\le \mathbb{P}_f\left(A<-\epsilon_2 t\right)+\mathbb{P}_f\left(B_1<-\epsilon_2 t\right)+\mathbb{P}_f\left(B_2<-\epsilon_2 t\right)+\left(B_3<-\epsilon_2 t\right)\\
&\le \exp\left(\tfrac{-\epsilon_2^2 t}{2\overline{\mathcal{L}}^2}\right)+\exp\left(\tfrac{-\epsilon_2^2 t}{2\overline{\mathcal{L}}^2}\right)+\mathbb{P}_f\left(\hat{\tau}\ge \tfrac{\epsilon_2 t}{2\overline{\mathcal{L}}}\right)+0 && \left[\text{due to (38)-(41); } t\ge\sqrt{2\overline{\mathcal{L}}/\epsilon_2}\right]\\
&=\mathbb{P}_f\left(\hat{\tau}\ge \tfrac{\epsilon_2 t}{2\overline{\mathcal{L}}}\right)+2\exp\left(\tfrac{-\epsilon_2^2 t}{2\overline{\mathcal{L}}^2}\right)=\tilde{\rho}_t
\end{aligned}
$$

Hence $\{\tilde{\rho}_t\}_t$ defined in Equation (43) satisfies Equation (37), and the proof is finished. ■

Appendix D: Proof of Theorem 3

D.1. Preliminaries

With a slight abuse of notation, let us first replace the optimization problem $\inf_{f\in\mathcal{M}_p}\sup_{\lambda\in\Delta(\mathcal{S})}\inf_{\bar{f}\in\overline{\mathcal{M}}_p(f)} D_\lambda\left(f||\bar{f}\right)$ with $\min_{f\in\mathcal{M}_p}\max_{\lambda\in\Delta(\mathcal{S})}\min_{\bar{f}\in\overline{\mathcal{M}}_p(f)} D_\lambda\left(f||\bar{f}\right)$ w.l.o.g based on the following observations: first, we study a relaxation of the original problem by taking the closures of $\mathcal{M}_p$ and $\overline{\mathcal{M}}_p(f)$ respectively, which are compact sets. In the relaxed problem, all the "inf" and "sup" can be replaced with "min" and "max" respectively because of the continuity of our objective function. Second, we may verify (ex-post) that our proposed solution (to the relaxed optimization problem) are interior points, i.e., feasible in the original optimization problem.

For a given ranking $\sigma\in\Sigma$, let us define $\mathcal{M}_p(\sigma)$ to be the subset of p-Separable preferences that are consistent with σ, that is, $\mathcal{M}_p(\sigma):=\{f\in\mathcal{M}_p\colon \sigma_f=\sigma\}$. We also define $\overline{\mathcal{M}}_p(\sigma):=\{f\in\mathcal{M}_p\colon \sigma_f(1)\neq\sigma(1)\}$. For $k\in[K]$, let $\Sigma_k:=\{\sigma\in\Sigma\colon \sigma(k)=1\}$ be the set of rankings that rank version k as the top-ranked version. For each k, we distinguish one ranking $\hat{\sigma}_k\in\Sigma_k$ that satisfies

$$
\hat{\sigma}_k(i)=\begin{cases} i+1 & \text{if } i=1,\dots,k-1\\ 1 & \text{if } i=k\\ i & \text{if } i=k+1,\dots,K.\end{cases} \tag{44}
$$

Recall (see footnote 7) that the preference f^{OA} can be identified up to permutations. So to simplify our notation, and without loss of generality, let us assume that the consensus ranking is equal to the identity ranking, i.e., $\sigma_*:=(1,2,\dots,K)$ and let us compute f^{OA} with respect to this ranking.

$$
\min_{f\in\mathcal{M}_p(\sigma_*)}\ \max_{\lambda\in\Delta(\mathcal{S})}\ \min_{\bar{f}\in\overline{\mathcal{M}}_p(\sigma_*)} D_\lambda\left(f||\bar{f}\right). \tag{45}
$$

D.2. Main Body of the Proof

In what follows, we will derive an explicit characterization of f^{OA} by mean of two results that we have stated as lemmas. Our first intermediate result establishes a dominance structure among the rankings in Σ_k.

Lemma 4. *For any $f^* \in \mathcal{M}_p(\sigma_*)$, $\sigma_k \in \Sigma_k$ and $f \in \mathcal{M}_p(\sigma_k)$ there exists a $\hat{f} \in \mathcal{M}_p(\hat{\sigma}_k)$ such that for any $\lambda \in \Delta(\mathcal{S})$*

$$D_\lambda\left(f^*||\hat{f}\right) \leq D_\lambda\left(f^*||f\right).$$

An important implication of Lemma 4 is that it allows us to simplify the inner minimization in the definition of f^{OA} in (45). Indeed, instead of minimizing over the set $\overline{\mathcal{M}}_p(\sigma_*)$ we can conduct this minimization over the much smaller set $\bigcup_{k=2}^K \mathcal{M}_p(\hat{\sigma}_k)$. In other words, the optimization problem (45) can be rewritten as

$$\min_{f\in\mathcal{M}_p(\sigma_*)} \max_{\lambda\in\Delta(\mathcal{S})} \min_{\bar{f}\in\bigcup_{k=2}^K \mathcal{M}_p(\hat{\sigma}_k)} D_\lambda\left(f||\bar{f}\right),$$

or equivalently

$$\min_{f\in\mathcal{M}_p(\sigma_*)} \max_{\lambda\in\Delta(\mathcal{S})} \min_{k\in\{2,\ldots,K\}} \min_{\bar{f}\in\mathcal{M}_p(\hat{\sigma}_k)} D_\lambda\left(f||\bar{f}\right). \tag{46}$$

The reason to include explicitly the minimization over k is motivated by our next result.

Lemma 5. *For any $k \in \{2,\ldots,K\}$ and any $S \in \mathcal{S}$ the optimization problem*

$$\min_{f\in\mathcal{M}_p(\sigma_*)} \min_{\bar{f}\in\mathcal{M}_p(\hat{\sigma}_k)} D_S\left(f||\bar{f}\right) \tag{47}$$

admits the following solution:

$$f^*(X|S) = f^{\mathrm{OA}}_{\sigma_*}(X|S) \quad \textit{and} \quad \bar{f}^*(X|S) = f^{\mathrm{OA}}_{\hat{\sigma}_k}(X|S) \qquad X \in S.$$

We will show that $f^{\mathrm{OA}}_{\sigma_*}$ in Lemma 5, which is independent of both k and S, also solves (the outer maximization of) (45). We do so by coming up with variations of (45) below:

$$\begin{aligned}
\underline{L} &:= \max_{\lambda\in\Delta(\mathcal{S})} \min_{k\in\{2,\ldots,K\}} D_\lambda\left(f^{\mathrm{OA}}_{\sigma_*}||f^{\mathrm{OA}}_{\hat{\sigma}_k}\right) \\
&= \max_{\lambda\in\Delta(\mathcal{S})} \min_{k\in\{2,\ldots,K\}} \min_{f\in\mathcal{M}_p(\sigma_*)} \min_{\bar{f}\in\mathcal{M}_p(\hat{\sigma}_k)} D_\lambda\left(f||\bar{f}\right) && \text{[Lemma 5]} \\
&= \max_{\lambda\in\Delta(\mathcal{S})} \min_{f\in\mathcal{M}_p(\sigma_*)} \min_{k\in\{2,\ldots,K\}} \min_{\bar{f}\in\mathcal{M}_p(\hat{\sigma}_k)} D_\lambda\left(f||\bar{f}\right) && \text{[swapping minimization]} \\
&\leq \min_{f\in\mathcal{M}_p(\sigma_*)} \max_{\lambda\in\Delta(\mathcal{S})} \min_{k\in\{2,\ldots,K\}} \min_{\bar{f}\in\mathcal{M}_p(\hat{\sigma}_k)} D_\lambda\left(f||\bar{f}\right) && \text{[Max-Min inequality; see (a) below]} \\
&\leq \min_{f\in\mathcal{M}_p(\sigma_*)} \min_{k\in\{2,\ldots,K\}} \min_{\bar{f}\in\mathcal{M}_p(\hat{\sigma}_k)} \max_{\lambda\in\Delta(\mathcal{S})} D_\lambda\left(f||\bar{f}\right) && \text{[Max-Min inequality]} \\
&= \min_{k\in\{2,\ldots,K\}} \min_{f\in\mathcal{M}_p(\sigma_*)} \min_{\bar{f}\in\mathcal{M}_p(\hat{\sigma}_k)} \max_{\lambda\in\Delta(\mathcal{S})} D_\lambda\left(f||\bar{f}\right) && \text{[swapping minimization]} \\
&= \min_{k\in\{2,\ldots,K\}} \max_{\lambda\in\Delta(\mathcal{S})} \min_{f\in\mathcal{M}_p(\sigma_*)} \min_{\bar{f}\in\mathcal{M}_p(\hat{\sigma}_k)} D_\lambda\left(f||\bar{f}\right) && \text{[see (b) below]} \\
&= \min_{k\in\{2,\ldots,K\}} \max_{\lambda\in\Delta(\mathcal{S})} D_\lambda\left(f^{\mathrm{OA}}_{\sigma_*}||f^{\mathrm{OA}}_{\hat{\sigma}_k}\right) =: \overline{U} && \text{[Lemma 5]}
\end{aligned}$$

In the derivations above, note that (a) is a restatement of Problem (46), which is equivalent to Problem (45). Part (b) is because of the following two facts: (i) the function $D_\lambda\left(f||\bar{f}\right)$ is jointly convex in $(f,\bar{f})$ and linear in λ; and (ii) the domains $\mathcal{M}_p(\sigma_*)$ and $\mathcal{M}_p(\hat{\sigma}_k)$ are both convex. Finally, it is easy to see that $\underline{L} = \overline{U}$ because through standard arguments in finite linear programming. Hence all of the optimization problems above are equivalent. ■

D.3. Proof of Auxiliary Lemmas

Proof of Lemma 4. For any $S \in \mathcal{S}$, we will show that $D_S\left(f^*||\hat{f}\right) \leq D_S\left(f^*||f\right)$. Let $S = \{X_1, X_2, \ldots, X_s\}$, and let us label the versions in S such that $\sigma_*(X_i) < \sigma_*(X_{i+1})$. Also, for notational convenience, let $f_i^* = f^*(X_i|S)$ and $f_i = f(X_i|S)$ for $i = 1, \ldots, s$. Note that because of our labeling, we have $f_1^* \geq f_2^* \geq \cdots \geq f_s^*$. On the other hand, let $\sigma_k^{-1}(\cdot|S)$ be the inverse of the restriction of σ to S and define the permutation $\{j_i\}_{i=1}^s$ of the element in $[s]$ in such a way that $\sigma_k^{-1}(i|S) = X_{j_i}$. It follows that $f_{j_1} \geq f_{j_2} \geq \cdots \geq f_{j_s}$.

Let us define the preference $\hat{f}$ in the lemma. We identify two cases:

(i) If $k \in S$ then $\hat{f}_{j_1} = f_{j_1}$, $\hat{f}_i = f_{j_{i+1}}$ for $i < j_1$ and $\hat{f}_i = f_{j_i}$ for $i > j_1$.

(ii) If $k \notin S$ then $\hat{f}_i = f_{j_i}$ for $i = 1, \ldots, s$.

It is not hard to see that this definition of $\hat{f}$ is consistent with the requirement $\hat{f} \in \mathcal{M}_p(\hat{\sigma}_k)$. Now, the condition $D_S\left(f^*||\hat{f}\right) \leq D_S\left(f^*||f\right)$ is equivalent to

$$\sum_{i=1}^{s} f_i^* \ln(\hat{f}_i) \geq \sum_{i=1}^{s} f_i^* \ln(f_i).$$

This inequality follows from a straightforward application of the rearrangement theorem and the following three facts: (1) the sequence $\{f_i^*\}_{i=1}^s$ is nonincreasing in i; (2) the sequence $\{\ln(\hat{f}_i)\}_{i=1}^s$ is a rearrangement of the sequence $\{\ln(f_i)\}_{i=1}^s$; and (3) the sequence $\{\ln(\hat{f}_i)\}_{i=1}^s$ is either nonincreasing in i if $k \notin S$ or is nonincreasing in i after excluding the j_1^{th} term at which the two sequences coincide if $k \in S$ (since $\hat{f}_{j_1} = f_{j_1}$ in this case). ■

Proof of Lemma 5. If $k \notin S$ then by the definition of $\hat{\sigma}_k$ it follows that $\sigma_*(X|S) = \hat{\sigma}_k(X|S)$ for all $X \in S$. Thus, the proposed solution in the lemma satisfies $f(X|S) = \bar{f}(X|S)$ for all $X \in S$ and so $D_S\left(f||\bar{f}\right) = 0$, which is trivially optimal since the KL divergence is always nonnegative.

Suppose now that $k \in S$. To ease notation, let us assume that the set $S = \{1, 2, \ldots, s\}$ with $k \leq s$. Define the permutation matrix $M \in \{0,1\}^{s \times s}$ such that $M(i, i+1) = 1$ for $i = 1, \ldots, k-1$, $M(k, 1) = 1$ and $M(i,i) = 1$ for $i = k+1, \ldots s$, with all other entries $M_{ij} = 0$. We let $x_i := f(X|S)$ and $y_i = \bar{f}_i(X|S)$ for all $i = 1, \ldots, s$. We note that the requirement $f \in \mathcal{M}_p(\sigma_*)$ implies that $x_{i+1} \leq p\, x_i$ for $i = 1, \ldots, s-1$. On the other hand, the requirement that $\bar{f} \in \mathcal{M}_p(\hat{\sigma}_k)$ implies there exists a probability vector $z \in \mathbb{R}^s$ such that $y = Mz$ and $z_{i+1} \leq p\, z_i$ for $i = 1, \ldots, s-1$. It follows that the optimization problem in (47) can be rewritten as the following convex optimization problem (convexity follows the fact that the KL divergence is a convex function on the pair of probability distributions (x, y)):

$$\begin{aligned} & \min_{x,z} \sum_{i=1}^{s} x_i \left[\ln(x_i) - \ln\Big(\sum_{j=1}^{s} M_{ij}\, z_j \Big) \right] \\ \text{subject to} \quad & x_{i+1} \le p\, x_i, \qquad i = 1, \ldots, s-1 \\ & z_{i+1} \le p\, z_i, \qquad i = 1, \ldots, s-1 \\ & \sum_i x_i = \sum_i z_i = 1 \\ & x, z \ge 0. \end{aligned} \tag{P}$$

With the convention that $\lambda_0 = \beta_0 = \lambda_s = \beta_s = 0$, we define the Lagrange function

$$\mathscr{L}(x,z;\zeta;\eta;\theta;\gamma) = \sum_{i=1}^{s} \left[x_i \ln(x_i) - x_i \ln\left(\sum_{j=1}^{s} M_{ij}\, z_j \right) + \zeta_i\,(x_{i+1} - p x_i) + \eta_i\,(z_{i+1} - p\, z_i) + \theta x_i + \gamma z_i \right].$$

Noticing that (P) is a convex optimization problem over linear constraints, the following KKT conditions are guarantees of optimality: for all $i = 1, \ldots, s$,[14]

$$\begin{aligned} \texttt{Stationarity in } x: \quad & \frac{\partial \mathscr{L}}{\partial x_i} = \ln(x_i) + 1 - \ln\left(\sum_{\ell=1}^{s} M_{i\ell}\, z_\ell \right) + \zeta_{i-1} - p\,\zeta_i + \theta = 0 \\ \texttt{Stationarity in } z: \quad & \frac{\partial \mathscr{L}}{\partial z_i} = -\sum_{\ell=1}^{s} x_\ell\, M_{\ell i}/z_i + \eta_{i-1} - p\,\eta_i + \gamma = 0 \\ \texttt{Complementary Slackness}: \quad & \zeta_i\,(x_{i+1} - p\, x_i) = 0 \quad \text{and} \quad \eta_i\,(z_{i+1} - p\, z_i) = 0 \\ \texttt{Primal Feasibility}: \quad & x_{i+1} \le p\, x_i, \quad z_{i+1} \le p\, z_i, \quad x_i \ge 0, \quad z_i \ge 0, \quad \text{and} \quad \sum_i x_i = \sum_i z_i = 1 \\ \texttt{Dual Feasibility}: \quad & \zeta_i \ge 0 \quad \text{and} \quad \eta_i \ge 0. \end{aligned}$$

Let us revert our change of variable and set $y^* = M\, z^*$ as well as $\tilde{y}^* = M^T\, x^*$. That is:

$$y_i^* = \begin{cases} z_{i+1}^* & \text{if } i = 1, \ldots, k-1 \\ z_1^* & \text{if } i = k \\ z_i^* & \text{if } i = k+1, \ldots, s \end{cases} \quad \text{and} \quad \tilde{y}_i^* = \begin{cases} x_i^* & \text{if } i = 1 \\ x_{i-1}^* & \text{if } i = 2, \ldots, k \\ x_i^* & \text{if } i = k+1, \ldots, s. \end{cases}$$

In addition, given $i \in [s]$, let us introduce $\Lambda_i = \sum_{\ell=1}^{i} p^{\ell-1} = (1-p^i)/(1-p)$ for shorthand notations. We postulate the following solution to the KKT conditions

$$\begin{aligned} & x_i^* = z_i^* = p^{i-1}/\Lambda_s \\ & \theta^* = -\left(\sum_{\ell=1}^{s} p^{\ell-1} \ln \tfrac{x_\ell^*}{y_\ell^*} \right) / \Lambda_s - 1 \\ & \zeta_i^* = \tfrac{1}{p^i} \left[\sum_{\ell=1}^{i} p^{\ell-1} \ln \tfrac{x_\ell^*}{y_\ell^*} + (1+\theta^*)\Lambda_i \right] \\ & \gamma^* = \left(\sum_{\ell=1}^{s} p^{\ell-1}\, \tfrac{\tilde{y}_\ell^*}{z_\ell^*} \right) / \Lambda_s \end{aligned}$$

[14] Implicitly, we are setting the dual variables for the constraint "$x, z \ge 0$" to zero.

$$\eta_i^* = \tfrac{1}{p^i}\left[\gamma^* \Lambda_i - \sum_{\ell=1}^{i} p^{\ell-1}\, \tfrac{\tilde{y}_\ell^*}{z_\ell^*}\right].$$

One can verify that, by construction, the $\{x_i^*\}$ satisfy $x_{i+1}^* = p\, x_i^*$ for all $i = 1, \dots, s-1$ and $\sum_i x_i^* = 1$; and the same is true for the $\{z_i^*\}$. Hence, `Complementary Slackness` and `Primal Feasibility` are directly satisfied. Similarly, it is not hard to see that the value of θ^*, ζ_i^*, γ^* and η_i^* are chosen so that both `Stationarity` conditions are satisfied. Hence, we only need to check `Dual Feasibility` for ζ_i and η_i.

Let us first verify that $\zeta_i^* \geq 0$. It is equivalent to

$$\frac{1}{\Lambda_i} \sum_{\ell=1}^{i} p^{\ell-1} \ln \frac{x_\ell^*}{y_\ell^*} \geq \frac{1}{\Lambda_s - \Lambda_i} \sum_{\ell=i+1}^{s} p^{\ell-1} \ln \frac{x_\ell^*}{y_\ell^*}. \tag{48}$$

Invoking the expressions for x^* and y^*, we notice that

$$p^{\ell-1} \ln \frac{x_\ell^*}{y_\ell^*} = \begin{cases} p^{\ell-1} \ln \frac{1}{p} & \text{if } \ell = 1, \dots, k-1 \\ -p^{k-1}(k-1) \ln \frac{1}{p} & \text{if } \ell = k \\ 0 & \text{if } \ell = k+1, \dots, s. \end{cases}$$

Thus for $i < k$ condition (48) becomes

$$\ln \frac{1}{p} \geq \frac{\ln \frac{1}{p}}{\Lambda_s - \Lambda_i} \left[p^i + \cdots + p^{k-2} - (k-1)p^{k-1}\right].$$

Noticing that the term of the left is positive and that on the right is negative, one can check this inequality holds. For $i \geq k$ condition (48) becomes

$$\frac{\ln \frac{1}{p}}{\Lambda_i} \left[1 + \cdots + p^{k-2} - (k-1)p^{k-1}\right] \geq 0,$$

which is also satisfied.

Let us then verify that $\eta_i^* \geq 0$. Invoking the expressions for $\tilde{y}^*$ and z^*, we notice that

$$p^{\ell-1} \frac{\tilde{y}_\ell^*}{z_\ell^*} = \begin{cases} p^{k-1} & \text{if } \ell = 1 \\ p^{\ell-2} & \text{if } \ell = 2, \dots, k \\ p^{\ell-1} & \text{if } \ell = k+1, \dots, s. \end{cases}$$

In particular, the vector $\left(p^{\ell-1} \frac{\tilde{y}_\ell^*}{z_\ell^*}\right)_{\ell=1,\dots,s}$ is a permutation of the vector $\left(p^{\ell-1}\right)_{\ell=1,\dots,s}$. As a result, $\gamma^* = 1$ and $\eta_i^* \geq 0$ is equivalent to

$$\sum_{\ell=1}^{i} p^{\ell-1} \geq \sum_{\ell=1}^{i} p^{\ell-1} \frac{\tilde{y}_\ell^*}{z_\ell^*}.$$

The inequality is satisfied because the vector $\left(p^{\ell-1}\right)_{\ell=1,\dots,s}$ is a decreasing sequence. ■

Appendix E: Proof of Theorem 4

E.1. Preliminaries

Let us introduce some notation. Let us introduce $d_\lambda(\sigma) := D_\lambda\left(f_{\sigma_*}^{\text{OA}} || f_\sigma^{\text{OA}}\right)$, for every $\sigma \in \Sigma$, a short-hand notation for the "distance" from ranking σ to the identity mapping σ_*. Invoking Corollary 1, a key step is to observe that (5) is equivalent to $\max_{\lambda \in \Delta(\mathcal{S})} \min_{\bar{f} \in \overline{\mathcal{M}}_p^{\text{OA}}(f)} D_\lambda\left(f || \bar{f}\right)$, which is further equivalent to the following LP:

$$\begin{aligned} \max_{\lambda, u} \quad & u \\ s.t. \quad & \sum_{S \in \mathcal{S}} d_S(\bar{\sigma}) \cdot \lambda(S) \geq u, \quad \forall \bar{\sigma} \in \overline{\Sigma}(\sigma_*) \\ & \sum_{S \in \mathcal{S}} \lambda(S) = 1 \\ & \lambda(S) \geq 0, \quad \forall S \in \mathcal{S} \end{aligned} \tag{LP-P}$$

In the expression above, we leverage that fact that each element $f \in \mathcal{M}_p^{\text{OA}}$ uniquely corresponds to a ranking $\sigma \in \Sigma$. Moreover, $\overline{\Sigma}(\sigma_*) := \{\sigma \in \Sigma : \sigma^{-1}(1) \neq 1\}$ is defined as the set of rankings which disagrees with σ_* in terms of the top-ranked item. We may also write out the dual problem of (LP-P) in the follows:

$$\begin{aligned} \min_{\mu, l} \quad & l \\ s.t. \quad & \sum_{\bar{\sigma} \in \overline{\Sigma}(\sigma_*)} d_S(\bar{\sigma}) \cdot \mu(\bar{\sigma}) \leq l, \quad \forall S \in \mathcal{S} \\ & \sum_{\bar{\sigma} \in \overline{\Sigma}(\sigma_*)} \mu(\bar{\sigma}) = 1 \\ & \mu(\bar{\sigma}) \geq 0, \quad \forall \bar{\sigma} \in \overline{\Sigma}(\sigma_*). \end{aligned} \tag{LP-D}$$

The dual problem is not used in the algorithm, but important in our analysis. We also introduce some other notation. Let us define, for $n \in [K]$, $s_n = (1-p)p^{n-1}$, so that the following equalities hold:

$$\begin{aligned} \sum_{k=1}^{n-1} (s_k - s_n) \log \frac{s_k}{s_{k+1}} &= \log\left(\frac{1}{p}\right)(1-p)\left[1 + p + \cdots + p^{n-2} - (n-1)p^{n-1}\right] = \mathfrak{a}_n \\ \sum_{k=1}^{n} s_k &= (1-p)(1 + p + \cdots + p^{n-1}) = 1 - p^n = \mathfrak{b}_n. \end{aligned} \tag{49}$$

Recall from the proof of Theorem 3 that $\hat{\sigma}_m := (2, 3, \ldots, m, 1, m+1, \ldots, K)$, for $m = 2, \ldots, K$. The introduction of $\mathfrak{a}_n$ and $\mathfrak{b}_n$ in (6) is helpful in evaluating $d_S(\hat{\sigma}_m)$, as stated in the lemma below. The proof of all technical lemmas in this subsection are in Section E.3.

Lemma 6. *Given any set $S \in \mathcal{S}$ and $m = 2, \ldots, K$,*

$$d_S(\hat{\sigma}_m) = \begin{cases} \mathfrak{a}_i / \mathfrak{b}_n & \text{if } m \in S \\ 0 & \text{if } m \notin S \end{cases}, \quad \text{where} \quad i = \sigma_*(m|S) \text{ and } n = |S|.$$

Ultimately we aim to solve both (LP-P) and (LP-D) in closed form. On one hand, $\lambda_*^{\mathrm{OA}}(\cdot)$ defined in (7) are closely related to the primal optimal solutions to (LP-P). On the other hand, we define the constants

$$\mu_m = \begin{cases} \frac{\mathfrak{b}_2}{\mathfrak{a}_2}, & \text{if } m = 2; \\ \frac{1}{\mathfrak{a}_m}(\mathfrak{b}_m - \mathfrak{b}_{m-1}), & \text{if } m = 3, \dots, K. \end{cases} \tag{50}$$

and

$$\mu^*(\bar{\sigma}) = \begin{cases} \frac{\mu_m}{\mu_2 + \dots + \mu_K}, & \text{if } \bar{\sigma} = \hat{\sigma}_m, \ m = 2, \dots, K; \\ 0, & \text{otherwise.} \end{cases} \tag{51}$$

The quantities $\{\mu^*(\bar{\sigma})\}_{\bar{\sigma}}$ are closely related to the dual optimal solutions to (LP-D).

In the lemmas that follow, we will show that there exist (u^*, v^*) such that $u^* = l^*$ and $(\lambda_*^{\mathrm{OA}}, u^*)$ and (μ^*, l^*) are primal and dual feasible in (LP-P) and (LP-D) respectively. This means both problems have the same objective value. By weak duality, this implies that $(\lambda_*^{\mathrm{OA}}, u^*)$ and (μ^*, l^*) are primal and dual optimal in (LP-P) and (LP-D) respectively.

Lemma 7. *$(\lambda_*^{\mathrm{OA}}, u^*)$ is feasible in (LP-P), where $u^* := \frac{1}{\lambda_2^* + \dots + \lambda_K^*}$.*

Lemma 8. *(μ^*, l^*) is feasible in (LP-D), where $l^* := \frac{1}{\mu_2 + \dots + \mu_K}$.*

Lemma 9. *$u^* = l^*$.*

Given any $S \in \mathcal{S}$, we define the reduced cost r of the optimal solution as

$$r(S) := \sum_{\sigma \in \overline{\Sigma}(\sigma_*)} d_S(\sigma) \cdot \mu^*(\sigma) - l^*. \tag{52}$$

By the proof of Lemma 8 below, we can also show the following corollary regarding the reduced costs, which guarantees the uniqueness of the optimal solution λ_*^{OA}.

Corollary 3. *For every S where $\lambda_*^{\mathrm{OA}}(S) = 0$, $r(S) < 0$.*

E.2. Main Body of Proof

Proof of Theorem 4. Because of the primal feasibility of $(\lambda_*^{\mathrm{OA}}, u^*)$ in Lemma 7, dual feasibility of (μ^*, l^*) in Lemma 8, and the objective values $u^* = l^*$ due to Lemma 9, we know that λ_*^{OA} is an optimal solution to (Max-Min) by weak duality (see Dantzig 1963). Due to Corollary 3, the reduced costs outside the support of λ_*^{OA} are strictly negative. Hence the optimal solution λ_*^{OA} is unique (see Dantzig 1963). ■

E.3. Proofs of Auxiliary Lemmas.

Proof of Lemma 6. The proof of this lemma is by calculation and verification. Fix an arbitrary $m \in \{2, \dots, K\}$ and $S = \{k_1, k_2, \dots, k_n\} \in \mathcal{S}$, where $1 \le k_1 < k_2 < \dots < k_n \le K$.

To calculate $d_S(\hat{\sigma}_m)$, it suffices to explicitly write out $f_{\sigma_*}^{\mathrm{OA}}$ and $f_{\hat{\sigma}_m}^{\mathrm{OA}}$, the p.m.f. of votes under ranking under σ_* and $\hat{\sigma}_m$, conditional on the display set being S. Recall that the restricted ranking is defined as $\sigma(k|S) := \sum_{i \in S} \mathbb{I}\{\sigma(i) \le \sigma(k)\}$, for all $\sigma \in \Sigma, S \in \mathcal{S}$ and $k \in S$. In particular, $\sigma_*(k_j|S) = j$

for all $j = 1, 2, \ldots, n$. Invoking the definition of f_σ^{OA} in (3), we can explicitly write out the p.m.f. of votes under ranking σ_* and display set S:

$$f_{\sigma_*}^{\mathrm{OA}}(k_j|S) = \frac{(1-p)p^{j-1}}{1-p^n} = \frac{s_j}{\mathfrak{b}_n}, \quad \forall j = 1, \ldots, n. \tag{53}$$

The closed form expression of $f_{\sigma_*}^{\mathrm{OA}}$ conditional on the display set being S, however, depends on the whether m is included in S:

Case 1: If $m \notin S$, S is a subset of $[K] \setminus \{m\}$. Hence $\hat{\sigma}_m(\cdot|S) = (1, 2, \cdots, n)$. Invoking (53),

$$f_{\hat{\sigma}_m}^{\mathrm{OA}}(k_j|S) = f_{\sigma_*}^{\mathrm{OA}}(k_j|S) = s_j/\mathfrak{b}_n, \quad \forall j = 1, \ldots, n. \tag{54}$$

That means $d_S(\hat{\sigma}_m) = \sum_{j=1}^n f_{\sigma_*}^{\mathrm{OA}}(k_j|S) \log \frac{f_{\sigma_*}^{\mathrm{OA}}(k_j|S)}{f_{\hat{\sigma}_m}^{\mathrm{OA}}(k_j|S)} = 0$.

Case 2: If $m \in S = \{k_1, k_2, \cdots, k_n\}$, there exists $i \in [n]$ so that $m = k_i$. Recall that $\sigma_*(k_j|S) = j$ for all $j = 1, 2, \ldots, n$. In particular, since $m = k_i$, $\sigma_*(m|S) = \sigma_*(k_i|S) = i$ Also, the restricted ranking $\hat{\sigma}_m(\cdot|S)$ is such that $\hat{\sigma}_m(k_i|S) = 1$, $\hat{\sigma}_m(k_j|S) = j+1$ for all $j = 1, \ldots, i-1$, and $\hat{\sigma}_m(k_j|S) = j$ for all $j = i+1, \ldots, n$. Invoking (3), we may write $f_{\hat{\sigma}_m}^{\mathrm{OA}}(\cdot|S)$ in closed form below:

$$f_{\hat{\sigma}_m}^{\mathrm{OA}}(k_j|S) = \begin{cases} s_{j+1}/\mathfrak{b}_n, & \text{if } j = 1, \ldots, i-1 \\ s_1/\mathfrak{b}_n, & \text{if } j = i \\ s_j/\mathfrak{b}_n, & \text{if } j = i+1, \ldots, n \end{cases} \tag{55}$$

That means

$$\begin{aligned} d_S(\hat{\sigma}_m) =& D_\lambda\left(f_{\sigma_*}^{\mathrm{OA}} || f_\sigma^{\mathrm{OA}}\right) \\ =& \sum_{j=1}^n f_{\sigma_*}^{\mathrm{OA}}(k_j|S) \log \frac{f_{\sigma_*}^{\mathrm{OA}}(k_j|S)}{f_{\hat{\sigma}_m}^{\mathrm{OA}}(k_j|S)} \\ =& \sum_{j=1}^{i-1} \frac{s_j}{\mathfrak{b}_n} \log \frac{s_j}{s_{j+1}} + \frac{s_i}{\mathfrak{b}_n} \log \frac{s_i}{s_1} + \sum_{j=i+1}^n \frac{s_j}{\mathfrak{b}_n} \log \frac{s_j}{s_j} && [(53) \text{ and } (55)] \\ =& \sum_{j=1}^{i-1} \frac{s_j}{\mathfrak{b}_n} \log \frac{s_j}{s_{j+1}} - \frac{s_i}{\mathfrak{b}_n} \left(\log \frac{s_1}{s_2} + \cdots + \log \frac{s_{i-1}}{s_i}\right) + 0 && [\text{expand } \log \frac{s_i}{s_1}] \\ =& \sum_{j=1}^{i-1} \frac{s_j - s_i}{\mathfrak{b}_n} \log \frac{s_j}{s_{j+1}} = \frac{\mathfrak{a}_i}{\mathfrak{b}_n}. && [(49)] \end{aligned}$$

By combining the two cases above, we finish the proof. ■

Proof of Lemma 7. Recall from (7) that we have defined λ_*^{OA} as:

$$\lambda_*^{\mathrm{OA}}(S) = \begin{cases} \frac{\lambda_n^*}{\lambda_2^* + \ldots + \lambda_K^*} & \text{if } S = [n] \text{ for some } n \in \{2, \ldots, K\} \\ 0 & \text{otherwise.} \end{cases}$$

To verify that λ_*^{OA} is primal feasible, we break the discussion into five steps. The first two steps check the first feasibility constraint in (LP-P), and the rest of the steps check the remaining feasibility constraints.

Step 1. We claim that $\sum_{S\in\mathcal{S}} d_S\left(\hat{\sigma}_m\right)\cdot\lambda_*^{\mathrm{OA}}(S)=u^*=\frac{1}{\lambda_2^*+\ldots+\lambda_K^*},\quad \forall m=2,\ldots,K.$

We know that λ_*^{OA} is only positive on S of form $[n]=\{1,\ldots,n\}$, where $n=2,\ldots,K$. Hence

$$\begin{aligned}
\sum_{S\in\mathcal{S}} d_S\left(\hat{\sigma}_m\right)\cdot\lambda_*^{\mathrm{OA}}(S)&=\sum_{n=2}^{K} d_{[n]}\left(\hat{\sigma}_m\right)\cdot\frac{\lambda_n^*}{\lambda_2^*+\cdots+\lambda_K^*} \qquad [\lambda_*^{\mathrm{OA}} \text{ defined in (7)}]\\
&\overset{(a)}{=}\sum_{n=m}^{K}\frac{\mathfrak{a}_m}{\mathfrak{b}_n}\cdot\frac{\lambda_n^*}{\lambda_2^*+\cdots+\lambda_K^*}\\
&=\frac{\mathfrak{a}_m}{\lambda_2^*+\cdots+\lambda_K^*}\left(\sum_{n=m}^{K}\frac{\lambda_n^*}{\mathfrak{b}_n}\right)\\
&\overset{(b)}{=}\frac{\mathfrak{a}_m}{\lambda_2^*+\cdots+\lambda_K^*}\cdot\left(\frac{1}{\mathfrak{a}_m}-\frac{1}{\mathfrak{a}_{m+1}}+\frac{1}{\mathfrak{a}_{m+1}}-\frac{1}{\mathfrak{a}_{m+2}}+\cdots+\frac{1}{\mathfrak{a}_K}\right)\\
&=\frac{1}{\lambda_2^*+\cdots+\lambda_K^*}.
\end{aligned}$$

In the chain of equalities, part (a) is by linking the fact that $\sigma_*\left(m|[n]\right)=m$ for $n\geq m$ to the calculation of $d_{[n]}\left(\hat{\sigma}_m\right)$ in Lemma 6. Part (b) is due to the definition of λ_n^* in (6).

Step 2. We claim that $\sum\limits_{S\in\mathcal{S}} d_S\left(\bar{\sigma}\right)\cdot\lambda_*^{\mathrm{OA}}(S)\geq u^*=\frac{1}{\lambda_2^*+\ldots+\lambda_K^*}$, for every $\bar{\sigma}\in\overline{\Sigma}(\sigma_*)$. Step 2 equivalent to the first line of constraints in Problem (LP-P).

Given any $\bar{\sigma}\in\overline{\Sigma}(\sigma_*)$, pick $m\in\{2,\ldots,K\}$ so that $\sigma(m)=1$. Invoking the dominance result in Lemma 4, we know that for every $S\in\mathcal{S}$, $d_S\left(\hat{\sigma}_m\right)\leq d_S\left(\bar{\sigma}\right)$. As a result, we have

$$\begin{aligned}
\sum_{S\in\mathcal{S}} d_S\left(\bar{\sigma}\right)\cdot\lambda_*^{\mathrm{OA}}(S)\geq&\sum_{S\in\mathcal{S}} d_S\left(\hat{\sigma}_m\right)\cdot\lambda_*^{\mathrm{OA}}(S)\\
=&\frac{1}{\lambda_2^*+\cdots+\lambda_K^*} \qquad [\text{due to Step 2}].
\end{aligned}$$

Step 3. We claim that $\sum_{S\in\mathcal{S}}\lambda_*^{\mathrm{OA}}(S)=1$. This step is trivial by (deliberate) construction of λ_*^{OA}. Step 3 is equivalent to the second line of constraints in Problem (LP-P).

Step 4. We claim that $\lambda_*^{\mathrm{OA}}(S)\geq 0$, for all $S\in\mathcal{S}$. Step 4 is equivalent to the last line of constraints in Problem (LP-P).

Due to the definition of λ_*^{OA} in (7), it suffices to verify that $\lambda_n^*\geq 0$, for every $n=2,\ldots,K$. Recall that $s_n=(1-p)p^{n-1}$ and λ_n^* is given by

$$\lambda_n^*=\begin{cases}\mathfrak{b}_n\left(\frac{1}{\mathfrak{a}_n}-\frac{1}{\mathfrak{a}_{n+1}}\right)=\frac{\mathfrak{b}_n}{\mathfrak{a}_n\mathfrak{a}_{n+1}}\cdot\left(\mathfrak{a}_{n+1}-\mathfrak{a}_n\right), & \text{if } n=2,\ldots,K-1;\\ \frac{\mathfrak{b}_K}{\mathfrak{a}_K}, & \text{if } n=K.\end{cases}$$

Notice the following three facts from (6):

1. $\mathfrak{b}_n=1-p^n>0$;
2. $\mathfrak{a}_n=\log\left(\frac{1}{p}\right)\left[1-np^{n-1}+(n-1)p^n\right]>0$;
3. $\mathfrak{a}_{n+1}-\mathfrak{a}_n=np^{n-1}(1-p)^2>0$.

Hence $\lambda_n^* \geq 0$, and the proof is complete. ■

Proof of Lemma 8. Recall from (51) that we have defined μ^* as:

$$\mu^*(\bar{\sigma}) = \begin{cases} \frac{\mu_m}{\mu_2+\ldots+\mu_K}, & \text{if } \bar{\sigma} = \hat{\sigma}_m, m = 2,\ldots,K; \\ 0, & \text{otherwise.} \end{cases}$$

We break the discussion into three steps. Each step corresponds to one line of constraints in Problem (LP-D) respectively.

Step 1. We claim that $\sum\limits_{\bar{\sigma}\in\overline{\Sigma}(\sigma_*)} d_S(\bar{\sigma}) \cdot \mu^*(\bar{\sigma}) \leq l^* = \frac{1}{\mu_2+\ldots+\mu_K}$, for every $S \in \mathcal{S}$. Step 1 is equivalent to the first line of constraints in Problem (LP-D).

Fix an arbitrary $S \in \mathcal{S}$, and suppose $|S| = n$. We know that $\mu^*(\cdot)$ is only positive on $\hat{\sigma}_m$, for $m = 2,\ldots,K$. As a result,

$$\begin{aligned}
&\sum_{\bar{\sigma}\in\overline{\Sigma}(\sigma_*)} d_S(\bar{\sigma}) \cdot \mu^*(\bar{\sigma}) \\
=&\sum_{m=2}^{K} d_S(\hat{\sigma}_m) \cdot \frac{\mu_m}{\mu_2+\cdots+\mu_K} && [\mu^* \text{ defined in } (51)] \\
=&\sum_{m=2}^{n} \frac{\mathfrak{a}_{\sigma_*(m|S)}}{\mathfrak{b}_n} \cdot \frac{\mu_m}{\mu_2+\cdots+\mu_K} && [\text{Lemma } 6] \\
\leq&\sum_{m=2}^{n} \frac{\mathfrak{a}_m}{\mathfrak{b}_n} \cdot \frac{\mu_m}{\mu_2+\cdots+\mu_K} && [\text{(i) } \sigma_*(m|S) \leq m; \text{ (ii) } \mathfrak{a}_n \uparrow \text{ in } n\ ] \\
=&\frac{1}{\mathfrak{b}_n(\mu_2+\cdots+\mu_K)} (\mathfrak{b}_2 + \mathfrak{b}_3 - \mathfrak{b}_2 + \cdots + \mathfrak{b}_n - \mathfrak{b}_{n-1}) && [\mu_m \text{ defined in } (50)] \\
=&\frac{1}{\mu_2+\cdots+\mu_K}.
\end{aligned}$$

Note that $\sigma_*(m|S) = m$ if and only if $[m] \subset S$, and $\mathfrak{a}_m$ is strictly increasing in m. Hence the inequality above becomes an equality if and only if $S = [n]$. This observation leads to Corollary 3 as a direct consequence.

Step 2. We claim that $\sum_{m=2}^{K} \mu^*(\hat{\sigma}_m) = 1$. This step is trivial by (deliberate) construction of μ^*. Step 2 is equivalent to the second line of constraints in Problem (LP-D).

Step 3. We claim that $\mu(\hat{\sigma}_m) > 0$, for all $m = 2,\ldots,K$. Step 3 is equivalent to the last line of constraints in Problem (LP-D).

For every $m \in \{2,\ldots,K\}$,

$$\mu(\hat{\sigma}_m) = \mu_m = \begin{cases} \frac{\mathfrak{b}_2}{\mathfrak{a}_2}, & \text{if } m = 2; \\ \frac{1}{\mathfrak{a}_m}(\mathfrak{b}_m - \mathfrak{b}_{m-1}), & \text{if } m = 3,\ldots,K \end{cases} = \begin{cases} \frac{s_1+s_2}{\mathfrak{a}_2}, & \text{if } m = 2; \\ \frac{s_m}{\mathfrak{a}_m}, & \text{if } m = 3,\ldots,K \end{cases} > 0.$$

The strict positiveness of μ_m is due to both the strict positiveness of $\mathfrak{a}_n$ and s_n, for $n, m \in \{2,\ldots,K\}$. That finishes the proof. ■

Proof of Corollary 3. This corollary is a restatement of the remark at the end of Step 1 of the proof of Lemma 8. ■

Proof of Lemma 9. Observe that

$$\begin{aligned}\sum_{n=2}^{K-1}\lambda_n^* &= \sum_{n=2}^{K-1}\mathfrak{b}_n\left(\frac{1}{\mathfrak{a}_n}-\frac{1}{\mathfrak{a}_{n+1}}\right)+\frac{\mathfrak{b}_K}{\mathfrak{a}_K}\\ &= \frac{\mathfrak{b}_2}{\mathfrak{a}_2}+\sum_{n=3}^{K}\frac{1}{\mathfrak{a}_n}(\mathfrak{b}_n-\mathfrak{b}_{n-1}) && \text{[summation by parts]}\\ &= \sum_{n=2}^{K-1}\mu_n && \text{[due to (50)].}\end{aligned}$$

As a result, $u^* = \frac{1}{\sum_{n=2}^{K-1}\lambda_n^*} = \frac{1}{\sum_{n=2}^{K-1}\mu_n} = l^*$. ■

Appendix F: Proofs of Proposition 2 and Corollary 2

Preliminaries. Throughout this section, the context of evaluating I_*^{OA} is clear. So let us suppress the argument and write $I_* = I_*^{\mathrm{OA}}$ for shorthand notation.

Proof of Proposition 2. We first show that $I_* = (1-p)\log\left(\frac{1}{p}\right)\left(1+\sum_{n=2}^{K}\frac{p^{n-1}}{1+2p+\cdots+(n-1)p^{n-2}}\right)^{-1}$.

Due to Lemma 7, 8, and 9, $I_* = \frac{1}{\lambda_2^*+\cdots+\lambda_K^*} = \frac{1}{\mu_2+\cdots+\mu_K} = \frac{1}{\frac{\mathfrak{b}_2}{\mathfrak{a}_2}+\sum_{n=3}^{K}\frac{1}{\mathfrak{a}_n}(\mathfrak{b}_n-\mathfrak{b}_{n-1})}$. Plug the values of $\mathfrak{a}_n, \mathfrak{b}_n$ in, and we have

$$\frac{1}{I_*} = \frac{\mathfrak{b}_2}{\mathfrak{a}_2}+\sum_{n=3}^{K}\frac{1}{\mathfrak{a}_n}(\mathfrak{b}_n-\mathfrak{b}_{n-1}) = \frac{\mathfrak{b}_2}{\mathfrak{a}_2}+\sum_{n=3}^{K}\frac{s_n}{\mathfrak{a}_n} = \frac{1}{\log\frac{1}{p}(1-p)}\left(1+p+\sum_{n=3}^{K}\frac{p^{n-1}}{1+2p+\cdots+(n-1)p^{n-2}}\right).$$

To give simple estimates for I_*, we break the rest of the discussion into two steps.

Step 1. We start with the upper bound. In fact, since $0<p<1$, $\sum_{n=3}^{K}\frac{p^{n-1}}{1+2p+\cdots+(n-1)p^{n-2}} \geq 0$. Hence

$$I_* = \frac{\log\frac{1}{p}(1-p)}{1+p+\sum_{n=3}^{K}\frac{p^{n-1}}{1+2p+\cdots+(n-1)p^{n-2}}} \leq \frac{\log\frac{1}{p}(1-p)}{1+p}.$$

Step 2. Next we establish the lower bound. Recall that we may rewrite I_* as $I_* = \frac{\Phi(p)}{\Psi(p)}$, where $\Phi(p) = \log\frac{1}{p}(1-p)$ and $\Psi(p) = 1+p+\sum_{n=3}^{K}\frac{p^{n-1}}{1+2p+\cdots+(n-1)p^{n-2}}$. Note that

$$\begin{aligned}\Psi(p) &= 1+p\left(\sum_{n=2}^{K}\frac{1}{\frac{1}{p^{n-2}}+\frac{2}{p^{n-3}}\cdots+(n-1)}\right) \leq 1+p\left(\sum_{n=2}^{K}\frac{1}{1+2+\cdots n-1}\right)\\ &= 1+p\left(\sum_{n=2}^{K}\frac{2}{n(n-1)}\right) = 1+2p\left(1-\frac{1}{2}+\frac{1}{2}-\frac{1}{3}+\cdots+\frac{1}{K-1}-\frac{1}{K}\right) = 1+2p\left(1-\frac{1}{K}\right)\end{aligned}$$

As a result, $I_* = \frac{\Phi(p)}{\Psi(p)} \geq \frac{\log\frac{1}{p}(1-p)}{1+2p(1-1/K)}$. That finishes the proof. ■

Proof of Corollary 2. Assume $K \geq 4$. We break the discussion into two steps.

Step 1. Choose an arbitrary $n \in \{2, \dots K-2\}$ We claim that $\lambda^*_{n+1} < \lambda^*_n$. We verify our claim by evaluating the term $\frac{\lambda^*_{n+1}}{\lambda^*_n}$ below:

$$\begin{aligned}
\frac{\lambda^*_{n+1}}{\lambda^*_n} &= \frac{\mathfrak{b}_{n+1}\frac{a_{n+2}-\mathfrak{a}_{n+1}}{a_{n+2}\mathfrak{a}_{n+1}}}{\mathfrak{b}_n\frac{a_{n+1}-\mathfrak{a}_n}{\mathfrak{a}_{n+1}\mathfrak{a}_n}} \\
&= \frac{\mathfrak{b}_{n+1}}{\mathfrak{b}_n}\,\frac{\mathfrak{a}_{n+2}-\mathfrak{a}_{n+1}}{\mathfrak{a}_{n+1}-\mathfrak{a}_n}\,\frac{\mathfrak{a}_n}{\mathfrak{a}_{n+2}} \\
&= \frac{1-p^{n+1}}{1-p^n}\;\frac{(n+1)p^n}{np^{n-1}}\;\frac{1-np^{n-1}+(n-1)p^n}{1-(n+2)p^{n+1}+(n+1)p^{n+2}} \\
&\overset{(a)}{=} \frac{(n+1)\big(p+p^2+\cdots+p^{n+1}\big)\big(1+2p+\cdots+(n-1)p^{n-2}\big)}{n\big(1+p+\cdots+p^{n-1}\big)\big(1+2p+\cdots+(n+1)p^n\big)}
\end{aligned}$$

Part (a) of the derivations above is due to the following two (algebraic) facts: For every $n \in \mathbb{Z}_+$,

$$\begin{aligned}
&\sum_{i=0}^{n-1} p^i = 1+2+\cdots+p^{n-1} = \tfrac{1-p^n}{1-p} \\
&\sum_{i=1}^{n} ip^{i-1} = 1+2p+\cdots+np^{n-1} = \tfrac{1+p+\cdots+p^{n-1}-np^n}{1-p} = \tfrac{1-(n+1)p^n+np^{n+1}}{(1-p)^2}
\end{aligned} \tag{56}$$

Note that $\frac{\lambda^*_{n+1}}{\lambda^*_n}$ is a ratio of two polynomials of p, namely, $\frac{\lambda^*_{n+1}}{\lambda^*_n} = \frac{(n+1)P(p)}{nQ(p)}$, where

$$\begin{aligned}
P(p) &:= \big(p+p^2+\cdots+p^{n+1}\big)\big(1+2p+\cdots+(n-1)p^{n-2}\big) \\
Q(p) &:= \big(1+p+\cdots+p^{n-1}\big)\big(1+2p+\cdots+(n+1)p^n\big).
\end{aligned}$$

Let $\{\xi_k\}_{k=1}^{2n-1}$ and $\{\nu_k\}_{k=1}^{2n-1}$ be the coefficients of $P(p)$ and $Q(p)$ respectively, so that $P(p) = \sum_{k=1}^{2n-1}\xi_k p^k$ and $Q(p) = \sum_{k=1}^{2n-1}\nu_k p^k + 1$. To show that $\lambda^*_{n+1} < \lambda^*_n$, it suffices to show that for every $k \in [2n-1]$, $(n+1)\xi_k < n\nu_k$ Observe that

$$\xi_k = \begin{cases} \frac{k(k+1)}{2}, & 1 \le k \le n-1 \\ \frac{n(n-1)}{2}, & k = n \\ \frac{(k-1)(2n-k)}{2}, & n+1 \le k \le 2n-1 \end{cases} \quad \text{and} \quad \nu_k = \begin{cases} \frac{(k+1)(k+2)}{2}, & 1 \le k \le n-1 \\ \frac{n(n+3)}{2}, & k = n \\ \frac{(k+3)(2n-k)}{2}, & n+1 \le k \le 2n-1 \end{cases}$$

Hence

$$\frac{(n+1)\xi_k}{n\nu_k} = \begin{cases} \frac{(n+1)k}{n(k+2)} \le \frac{(n+1)(n-1)}{n(n+1)} < 1, & 1 \le k \le n-1 \\ \frac{(n-1)(n+1)}{(n+3)n} = \frac{n^2-1}{n^2+3n} < 1, & k = n \\ \frac{(n+1)(k-1)}{n(k-3)} \le \frac{(n+1)(2n-2)}{n(2n-4)} = \frac{(n+1)(n-1)}{n(n-2)} < 1, & n+1 \le k \le 2n-1 \end{cases}$$

In conclusion, $\lambda^*_{n+1} < \lambda^*_n$ for every $2 \le n \le K-2$.

Step 2. We claim that $\lambda^*_{K-1} < \lambda^*_K$ (thus finishing the proof). We verify our claim by evaluating the term $\frac{\lambda^*_{K-1}}{\lambda^*_K}$ below:

$$\frac{\lambda^*_{K-1}}{\lambda^*_K} = \frac{\mathfrak{b}_{K-1}\frac{\mathfrak{a}_K-\mathfrak{a}_{K-1}}{\mathfrak{a}_{K-1}\mathfrak{a}_K}}{\mathfrak{b}_K/\mathfrak{a}_K} \qquad \text{[due to (6)]}$$

$$
\begin{aligned}
&= \frac{\mathfrak{b}_{K-1}}{\mathfrak{b}_K}\frac{\mathfrak{a}_K - \mathfrak{a}_{K-1}}{\mathfrak{a}_{K-1}} \\
&= \frac{1-p^{K-1}}{1-p^K}\frac{(K-1)p^{K-2}(1-p)^2}{1-(K-1)p^{K-2}+(K-2)p^{K-1}} \\
&= \frac{(K-1)p^{K-2}\left(1+p+\cdots+p^{K-2}\right)}{\left(1+p+\cdots+p^{K-1}\right)\left(1+2p+\cdots+(K-2)p^{K-3}\right)}. \qquad \text{[due to (56)]}
\end{aligned}
$$

Again, observe that $\frac{\lambda^*_{K-1}}{\lambda^*_K}$ is the ratio of two polynomials of p. Namely, $\frac{\lambda^*_{K-1}}{\lambda^*_K} = \frac{\tilde{P}(p)}{\tilde{Q}(p)}$, where

$$
\begin{aligned}
\tilde{P}(p) &:= (K-1)p^{K-2}\left(1+p+\cdots+p^{K-2}\right) \\
\tilde{Q}(p) &:= \left(1+p+\cdots+p^{K-1}\right)\left(1+2p+\cdots+(K-2)p^{K-3}\right)
\end{aligned}
$$

Denote $\{\tilde{\xi}_k\}_{k=0}^{2K-4}$ and $\{\tilde{\nu}_k\}_{k=0}^{2K-4}$ as the coefficients of $\tilde{P}(p)$ and $\tilde{Q}(p)$ to be respectively, so that $\tilde{P}(p) = \sum_{k=0}^{2K-4}\tilde{\xi}_k p^k$ and $\tilde{Q}(p) = \sum_{k=0}^{2K-4}\tilde{\nu}_k p^k$. To show $\lambda^*_{K-1} < \lambda^*_K$, it suffices to show that $\tilde{P}(p) < \tilde{Q}(p)$. By observation, one can see

$$
\tilde{\xi}_k = \begin{cases} 0, & 0 \le k \le K-3 \\ K-1, & K-2 \le k \le 2K-4 \end{cases} \quad \text{and} \quad \tilde{\nu}_k = \begin{cases} \frac{(k+1)(k+2)}{2}, & 0 \le k \le K-3 \\ \frac{(K-1)(K-2)}{2}, & k = K-2 \\ \frac{k(2K-3-k)}{2}, & K-1 \le k \le 2K-4 \end{cases}
$$

Hence

$$
\tilde{\nu}_k - \tilde{\xi}_k = \begin{cases} 1, & k = 0 \\ \frac{(k+1)(k+2)}{2} > 0, & 1 \le k \le K-3 \\ \frac{(K-1)(K-4)}{2} \ge 0, & k = K-2 \\ \frac{k(2K-3-k)}{2} - (K-1) \overset{(a)}{\ge} K-4 \ge 0, & K-1 \le k \le 2K-5 \\ -1, & k = 2K-4 \end{cases}
$$

In the derivations above, part (a) is by minimizing the term $\frac{k(2K-3-k)}{2} - (K-1)$ (as a quadratic function of k) subject to the constraint $K-1 \le k \le 2K-5$. This term obtains its minimal value at $K-4$ when $k = 2K-5$. Finally, we evaluate $\tilde{Q}(p) - \tilde{P}(p)$ below:

$$
\tilde{Q}(p) - \tilde{P}(p) = \sum_{k=0}^{2K-4}\tilde{\nu}_k p^k - \sum_{k=0}^{2K-4}\tilde{\nu}_k p^k = \sum_{k=0}^{2K-4}\left(\tilde{\nu}_k - \tilde{\xi}_k\right)p^K \ge 1 - p^{2K-4} > 0
$$

Hence $\lambda^*_{K-1} < \lambda^*_K$ and the proof is finished. ■

Appendix G: Proof of Theorem 6

The first part of the proof (i.e., proving (10)) is almost a repeat verbatim of that of Theorem 2 (Step 1 plus the corresponding Lemma 1), by taking $\mathcal{M}_p^F = \mathcal{M}_p^{OA}$ (which is a finite set) and taking $\beta^F = C_0 + \log(1/\delta)$ (see also Remark 4). There are, however, two differences to be mindful of.

1. In Step 3 of Algorithm 2, uniform randomization over all display sets is implemented when the time epoch t is a perfect square number. Under the Myopic Tracking Policy, we do not need such a component since $\lambda_*^{\mathrm{OA}}([K]) > 0$ by Theorem 4.

2. In the context of Theorem 6, the sample complexity guarantee (10) holds for an arbitrary $f \in \mathcal{M}_p$ rather than $f \in \mathcal{M}_p^{OA}$.

Because of the aforementioned differences, we need to present a slightly modified proof compared to that in Lemma 1. Without loss, suppose $f \in \mathcal{M}_p(\sigma_*)$. Pick $f^{\mathrm{OA}} = f^{\mathrm{OA}}_{\sigma_*}$ (see (3)) and an arbitrary $\bar{f}^{\mathrm{OA}} \in \overline{\mathcal{M}}^{\mathrm{OA}}$. Recall that $L^{f^{\mathrm{OA}},\bar{f}^{\mathrm{OA}}} : X, S \mapsto \log\left(\frac{f^{\mathrm{OA}}(X|S)}{\bar{f}^{\mathrm{OA}}(X|S)}\right)$ is the one-stage log-likelihood ratio function. Define

$$D_S^f\left(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}\right) := \mathbb{E}_{X\sim f}\left[L^{f^{\mathrm{OA}},\bar{f}^{\mathrm{OA}}}(X|S)\right].$$

This notation is consistent with that of Kullback-Leibler divergence because one can verify that $D_S^{f^{\mathrm{OA}}}(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}) = \mathbb{E}_{X\sim f^{\mathrm{OA}}}\left[L^{f^{\mathrm{OA}},\bar{f}^{\mathrm{OA}}}(X|S)\right] = D_S\left(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}\right)$.

Let us present a new version of Part 2 of the proof of Lemma 1, i.e., an estimate of the log likelihood ratio process $L_t^{f^{\mathrm{OA}},\bar{f}^{\mathrm{OA}}}$

$$\begin{aligned}
L_t^{f^{\mathrm{OA}},\bar{f}^{\mathrm{OA}}} &= \sum_{\ell=1}^{t} L^{f^{\mathrm{OA}},\bar{f}^{\mathrm{OA}}}(X_\ell, S_\ell) \\
&= \underbrace{\sum_{\ell=1}^{t}\left[L^{f^{\mathrm{OA}},\bar{f}^{\mathrm{OA}}}(X_\ell,S_\ell) - D_{S_\ell}^f\left(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}\right)\right]}_{A} + \underbrace{\sum_{\ell=1}^{t}\left[D_{S_\ell}^f\left(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}\right) - D^f_{\lambda_*^{\mathrm{OA}}}\left(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}\right)\right]}_{B} \\
&\quad + \underbrace{t \cdot D^f_{\lambda_*^{\mathrm{OA}}}\left(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}\right)}_{C}.
\end{aligned}$$

In the expression above, Part A corresponds to randomness from the choices (given the display sets). It is a $\mathbb{P}_f-$martingale with bounded differences (and hence diverges sublinearly). Part B corresponds to the randomness from display sets (since the algorithm decides which display set to offer at each epoch based on historical data and possible randomization). Part C corresponds to the deterministic part, i.e., the long-run growth rate of the process $L_t^{f^{\mathrm{OA}},\bar{f}^{\mathrm{OA}}}$. Similar to Part 2 of the proof of Lemma 1, we will show that both Part A and B diverge sublinearly while Part C grows at least as fast as $I_*^{\mathrm{OA}} t$.

Our key observation is that there exists $\hat{f}^{\mathrm{OA}} \in \overline{\mathcal{M}}_p^{\mathrm{OA}}$ such that for every display set $S \in \mathcal{S}$, $D_S^f\left(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}\right) \geq D_S\left(f^{\mathrm{OA}}||\hat{f}^{\mathrm{OA}}\right)$. In order to see that, let us suppose $S = [s]$ without loss of generality. Let $k := \sigma^{-1}_{\bar{f}^{\mathrm{OA}}}(1)$ be the top ranked item under preference $\bar{f}^{\mathrm{OA}}$, and $\hat{f}^{\mathrm{OA}} = f^{\mathrm{OA}}_{\hat{\sigma}_k}$, where $\hat{\sigma}_k$ is defined in (44). In addition, let us denote $f_i = f(X_i|S), f_i^{\mathrm{OA}} = f^{\mathrm{OA}}(X_i|S), \bar{f}_i^{\mathrm{OA}} = \bar{f}^{\mathrm{OA}}(X_i|S)$, and $\hat{f}_i^{\mathrm{OA}} = \hat{f}^{\mathrm{OA}}(X_i|S)$ for shorthand notations. Notice that

$$\begin{aligned}
&D_S^f\left(f^{\mathrm{OA}}||\bar{f}^{\mathrm{OA}}\right) \\
= &\sum_{i=1}^{s} f_i\left[\log\left(f_i^{\mathrm{OA}}\right) - \log\left(\bar{f}_i^{\mathrm{OA}}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \sum_{i=1}^{s} f_i \left[\log\left(f_i^{\mathrm{OA}}\right) - \log\left(\hat{f}_i^{\mathrm{OA}}\right)\right] \\
&\overset{(b)}{\geq} \log(1/p)\left(\sum_{i=1}^{k-1} f_i\right) + (k-1)\log(p) f_k \\
&\overset{(c)}{\geq} \log(1/p)\,\Lambda_{k-1}/\Lambda_s + (k-1)\,\log(p)\,p^{k-1}/\Lambda_s \qquad\qquad [\Lambda_n = 1+\cdots+p^{n-1} = \tfrac{1-p^s}{1-p}] \\
&\overset{(d)}{=} \sum_{i=1}^{k-1} \log(1/p) f_i^{\mathrm{OA}} + (k-1)\log(p) f_k^{\mathrm{OA}} \\
&= \sum_{i=1}^{s} f_i^{\mathrm{OA}} \left[\log\left(f_i^{\mathrm{OA}}\right) - \log\left(\hat{f}_i^{\mathrm{OA}}\right)\right] = D_S\left(f^{\mathrm{OA}} \| \hat{f}^{\mathrm{OA}}\right).
\end{aligned}
$$

In the derivations above, part (a) is a result of the rearrangement theorem plus the following three facts: (i) the sequence $\{f_i\}_{i=1}^s$ is decreasing in i; (ii) the sequence $\{\hat{f}_i^{\mathrm{OA}}\}_{i=1}^s$ is an rearrangement of the sequence $\{\bar{f}_i^{\mathrm{OA}}\}_{i=1}^s$; and (iii) the sequence $\{\hat{f}_i^{\mathrm{OA}}\}_{i=1}^s$ is either decreasing in i if $k \notin S$ or is nonincreasing in i after excluding the k^{th} term at which the two sequences $\{\hat{f}_i^{\mathrm{OA}}\}_{i=1}^s$ and $\{\bar{f}_i^{\mathrm{OA}}\}_{i=1}^s$ coincide if $k \in S$. Part (b) is by invoking the explicit expressions for f^{OA} and $\hat{f}^{\mathrm{OA}}$ respectively. Part (c) is a result of the fact that $f \in \mathcal{M}_p(\sigma_*)$ plus the following chain of expressions:

$$
\begin{aligned}
\sum_{i=1}^{k-1} f_i - (k-1) f_k &= \underbrace{\left(\sum_{i=1}^{k} f_i\right)}_{\geq \Lambda_k/\Lambda_s} \underbrace{\left(\frac{\sum_{i=1}^{k-1} f_i - (k-1) f_k}{\sum_{i=1}^{k} f_i}\right)}_{>0} \\
&\geq \frac{\Lambda_k}{\Lambda_s}\left(1 - k\,\underbrace{f_k/(f_1+\cdots+f_k)}_{\leq p^{k-1}/\Lambda_k}\right) \\
&\geq \frac{\Lambda_k}{\Lambda_s}\left(1 - k\,p^{k-1}/\Lambda_k\right) \\
&= \Lambda_{k-1}/\Lambda_s - (k-1)p^{k-1}/\Lambda_s.
\end{aligned}
$$

Part (d) is due to invoking the explicit expressions for f^{OA} again.

Our key observation has several implications. For example, the full display set strictly separates f^{OA} from any $\tilde{f}^{\mathrm{OA}} \in \mathcal{M}_p^{\mathrm{OA}} \setminus \{f^{\mathrm{OA}}\}$, i.e., $D_{[K]}^f\left(f^{\mathrm{OA}} \| \tilde{f}^{\mathrm{OA}}\right) \geq D_{[K]}\left(f^{\mathrm{OA}} \| \tilde{f}^{\mathrm{OA}}\right) > 0$. As a result, f_t^{OA} converges quickly to f^{OA} in probabilityt, i.e., $\mathbb{P}_f(f_t^{\mathrm{OA}} \neq f^{\mathrm{OA}}) \leq Ce^{-\epsilon t}$ for some $C, \epsilon > 0$ independent of δ.[15] A further consequence is that Part B grows sublinearly. To see why we can decomposing Part B into two parts: $B = \underbrace{\sum_{\ell=1}^{t}\left[D_{\tilde{S}_\ell}^f\left(f^{\mathrm{OA}} \| \bar{f}^{\mathrm{OA}}\right) - D_{\lambda_*^{\mathrm{OA}}}^f\left(f^{\mathrm{OA}} \| \bar{f}^{\mathrm{OA}}\right)\right]}_{B_1} + \underbrace{\sum_{\ell=1}^{t \wedge \hat{\tau}}\left[D_{S_\ell}^f\left(f^{\mathrm{OA}} \| \bar{f}^{\mathrm{OA}}\right) - D_{\lambda_*^{\mathrm{OA}}}^f\left(f^{\mathrm{OA}} \| \bar{f}^{\mathrm{OA}}\right)\right]}_{B_2}$, where $\hat{\tau} :=$

[15] Note that here the tail probability is improved from $Ce^{-\epsilon\sqrt{t}}$ in Lemma 3. The reason is that MTP displays a strictly separating set (i.e., $[K]$) with strictly positive probability at each time epoch. It is unclear whether this is true in the general setting. Instead, the proof relies on "forced exploration" when the time epoch is a perfect square number;see Step 3 of Algorithm 2.

$\max\{t : f_t^{\mathrm{OA}} \neq f^{\mathrm{OA}}\}$ and $\{\tilde{S}_\ell\}_\ell$ is a sequence of i.i.d. $\mathcal{S}$-valued random variables with distribution λ_*^{OA} such that $\tilde{S}_\ell = S_\ell$ for all $t \geq \ell \geq \hat{\tau}+1$. Here B_1 is a partial sum of i.i.d. random variables with mean zero, and B_2 takes a diminishing fraction of time epochs when t is large. Hence both diverge sublinearly. Finally, notice that

$$C = t \cdot D^f_{\lambda_*^{\mathrm{OA}}}\left(f^{\mathrm{OA}} || \bar{f}^{\mathrm{OA}}\right) \geq t \cdot D_{\lambda_*^{\mathrm{OA}}}\left(f^{\mathrm{OA}} || \hat{f}^{\mathrm{OA}}\right) \geq \min_{\bar{f} \in \overline{\mathcal{M}}_p^{\mathrm{OA}}} D_{\lambda_*^{\mathrm{OA}}}\left(f^{\mathrm{OA}} || \bar{f}\right) = t \cdot I_*^{\mathrm{OA}}.$$

The rest of the proof of (10) follows verbatim from the arguments in Step 1 of that of Theorem 2.

Finally, the second part of the proof, i.e., showing that (10) can be taken to be equality when $f \in \mathcal{M}_p^{OA}$ is a straightforward corollary of the lower bound result 1. ■

Appendix H: Proof of Theorem 7

Before we prove Theorem 7, let us first formally state our estimate of $\mathbb{P}_f(d_\tau \neq \sigma_f^{-1}(1))$ for an arbitrary preference $f \in \mathcal{M}_p$ in Lemma 10. Some notation is useful in the development of Lemma 10 below. Recall from (MLE) that f_t^{OA} is the most likely consensus preference at time t. For simplicity of notation, let us denote $\sigma_t := \sigma_{f_t^{\mathrm{OA}}}$ to be ranking associated with f_t^{OA}. For every given ranking σ, let us denote

$$\mathcal{E}_\sigma := \{\sigma_\tau = \sigma\} = \cup_{t=0}^{\infty} \{\sigma_t = \sigma\} \cap \{\tau = t\} \tag{57}$$

the event that the aggregated ranking equals σ when the algorithm terminates.

LEMMA 10. *There exists a constant $\tilde{C}_1$ (only dependent on K) such that for every preference instance $f \in \mathcal{M}_p$, $\mathbb{P}_f(d_\tau \neq \sigma_f^{-1}(1)) \leq \tilde{C}_1 e^{-\beta}$.*

Proof. Fix an arbitrary $f \in \mathcal{M}_p$ and $\bar{\sigma} \in \overline{\Sigma}(\sigma_f)$. We will show that the probability that the algorithm stops with the best-estimated ranking to be (mistakenly) $\bar{\sigma}$ is small, or $\mathbb{E}_f[\mathbb{I}\{\mathcal{E}_{\bar{\sigma}}\}] \leq e^{-\beta}$. That leads to our desired result because $\mathbb{P}_f(d_\tau \neq \sigma_f^{-1}(1)) \leq \sum_{\bar{\sigma} \in \overline{\Sigma}(\sigma_f)} \mathbb{P}_f(\mathcal{E}_{\bar{\sigma}}) \leq |\overline{\Sigma}(\sigma_f)| e^{-\beta}$. That finished the proof by letting $\tilde{C}_1 := |\overline{\Sigma}(\sigma_f)| = (K-1)(K-1)!$, where $n!$ represents the factorial of positive integer n.

For ease of notation, let us properly relabel the versions (within this proof only) so that $\bar{\sigma} = \sigma_*$. Under this relabeling rule, let $k := \sigma_f^{-1}(1)$, the top-ranked item under σ. We recall from (44) that $\hat{\sigma}_k$ is a particular ranking such that $\hat{\sigma}_k^{-1}(1) = \sigma_f^{-1}(1) = k$. We also pick $\bar{f}^{\mathrm{OA}} = f_{\sigma_*}^{\mathrm{OA}}, \hat{f}^{\mathrm{OA}} = f_{\hat{\sigma}_k}^{\mathrm{OA}}$, the expressions of which can be found in (3). Finally, for simplicity of notation, recall $\Lambda_n = 1 + p + \cdots + p^{n-1} = (1-p^n)/(1-p)$ for every integer n. Notice that

$$\begin{aligned}
\mathbb{E}_f[\mathbb{I}\{\mathcal{E}_{\bar{\sigma}}\}] =& \mathbb{E}_{\bar{f}^{\mathrm{OA}}}\left[\mathbb{I}\{\mathcal{E}_{\bar{\sigma}}\} \exp\left(-L_\tau^{\bar{f}^{\mathrm{OA}}, f}\right)\right] && \text{[change-of-measure]} \\
=& \mathbb{E}_{\bar{f}^{\mathrm{OA}}}\left[\mathbb{I}\{\mathcal{E}_{\bar{\sigma}}\} \exp\left(-L_\tau^{\bar{f}^{\mathrm{OA}}, \hat{f}^{\mathrm{OA}}}\right) \exp\left(-L_\tau^{\hat{f}^{\mathrm{OA}}, f}\right)\right] \\
\leq& \mathbb{E}_{\bar{f}^{\mathrm{OA}}}\left[e^{-\beta} \exp\left(-L_\tau^{\hat{f}^{\mathrm{OA}}, f}\right)\right] && \left[\mathbb{I}\{\mathcal{E}_{\bar{\sigma}}\} \exp\left(-L_\tau^{\bar{f}^{\mathrm{OA}}, \hat{f}^{\mathrm{OA}}}\right) \leq e^{-\beta} \text{ a.s.}\right]
\end{aligned}$$

$$
\begin{aligned}
=& e^{-\beta}\mathbb{E}_{\bar{f}^{\mathrm{OA}}}\left[\exp\left(-L_\tau^{\hat{f}^{\mathrm{OA}},f}\right)\right]\\
=& e^{-\beta}\mathbb{E}_{\bar{f}^{\mathrm{OA}}}\left[\tfrac{f(X_1|S_1)}{\hat{f}^{\mathrm{OA}}(X_1|S_1)}\tfrac{f(X_2|S_2)}{\hat{f}^{\mathrm{OA}}(X_2|S_2)}\cdots\tfrac{f(X_\tau|S_\tau)}{\hat{f}^{\mathrm{OA}}(X_\tau|S_\tau)}\right].
\end{aligned}
$$

Since X_t is independent of the history conditional on S_t, it suffices to show that for every display set $S\in\mathcal{S}$, $\mathbb{E}_{\bar{f}^{\mathrm{OA}}}\left[\frac{f(X|S)}{\hat{f}^{\mathrm{OA}}(X|S)}\right]\le 1$, where the expectation is taken over X only. To verify this inequality, first notice that if $k\notin S$, $f(X|S)\equiv\hat{f}^{\mathrm{OA}}(X|S)$ and $\frac{f(X|S)}{\hat{f}^{\mathrm{OA}}(X|S)}=1$ almost surely. Otherwise, it is without loss of generality to assume that $S=[s]$. Let us introduce the notation $f_i:=f(X_i|[K])$ for all $i\in[K]$. In that case,

$$
\begin{aligned}
\mathbb{E}_{\bar{f}^{OA}}\left[\tfrac{f(X|S)}{\hat{f}^{\mathrm{OA}}(X|S)}\right]-1=&\sum_{X\in[K]}\bar{f}^{OA}(X|[K])\tfrac{f(X|[K])}{\hat{f}^{\mathrm{OA}}(X|[K])}-1\\
=&\sum_{i=1}^{k-1}\tfrac{p^{i-1}}{\Lambda_K}\tfrac{f_i}{\frac{p^i}{\Lambda_K}}+\tfrac{p^{k-1}}{\Lambda_K}\tfrac{f_k}{\frac{1}{\Lambda_K}}+\sum_{i=k+1}^{K}\tfrac{p^{i-1}}{\Lambda_K}\tfrac{f_i}{\frac{p^{i-1}}{\Lambda_K}}-1\\
=&\sum_{i=1}^{k-1}\frac{1}{p}f_i+p^{k-1}f_k+\sum_{i=k+1}^{K}f_i-1\\
=&\left(\tfrac{1}{p}-1\right)\sum_{i=1}^{k-1}f_i+(p^{k-1}-1)f_k\\
\le&\left(\tfrac{1}{p}-1\right)\tfrac{p+p^2+\cdots+p^{k-1}}{\Lambda_K}+(p^{k-1}-1)\tfrac{1}{\Lambda_K}\\
=&\tfrac{1+\cdots+p^{k-1}-1-\cdots-p^{k-1}}{\Lambda_K}=0.
\end{aligned}
$$

That concludes the proof. ■

Proof of Theorem 7. The fact that $\mathbb{P}_f(\tau<\infty)=1$ for every $f\in\mathcal{M}_p$ is ensured by the proof of Theorem 6 Due to Lemma 10, we only need to have $\beta\ge\log(\tilde{C}_1)+\log\frac{1}{\delta}$ to ensure $\mathbb{P}_f(d_\tau\neq\sigma_f^{-1}(1))\le\delta$. Hence the proof is finished by letting $C_1=\log(\tilde{C}_1)$. ■

Appendix I: Proof of Propositions 3, 4, 5 and 6

Proof of Proposition 3. We first show part 1. Due to Theorem 3, we have

$$
f_\sigma^{\mathrm{OA}}(X|S)=\frac{1-p}{(1-p^{|S|})}p^{\sigma(k|S)-1},\quad\forall\sigma\in\Sigma,S\in\mathcal{S},k\in S.
$$

In the above expression, the relative ranking is defined as $\sigma(k|S):=\sum_{i\in S}\mathbb{I}\{\sigma(i)\le\sigma(k)\}$. For every $f\in\mathcal{M}_p^{OA}$ there exists a σ so that $f=f_\sigma^{\mathrm{OA}}$. For all voting history $H_t=(S_1,X_1,\dots,S_t,X_t)$,

$$
\begin{aligned}
\sum_{\ell=1}^{t}\log f_\sigma^{\mathrm{OA}}(X_\ell|S_\ell)=&\sum_{\ell=1}^{t}\left[\log\left(\frac{1-p}{p\,(1-p^{|S_\ell|})}\right)+\log p\cdot\left(\sum_{i\in S_\ell}\sigma(i)\le\sigma(X_\ell)\right)\right]\\
=&\sum_{\ell=1}^{t}\log\left(\tfrac{1-p}{1-p^{|S_\ell|}}\right)+\log p\cdot\left(\sum_{\ell=1}^{t}\sum_{(i,j):i\neq j}\mathbb{I}\{\sigma(j)<\sigma(i)\}\cdot\mathbb{I}\{X_\ell=i\}\cdot\mathbb{I}\{i,j\in S_\ell\}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{\ell=1}^{t} \log\left(\tfrac{1-p}{1-|p|^{S_\ell}}\right) + \log p \cdot \sum_{(i,j):i\neq j} \left[\mathbb{I}\{\sigma(j) < \sigma(i)\} \cdot \left(\sum_{\ell=1}^{t} \mathbb{I}\{X_\ell = i\} \cdot \mathbb{I}\{i,j \in S_\ell\} \right) \right] \\
&= \sum_{\ell=1}^{t} \log\left(\tfrac{1-p}{1-|p|^{S_\ell}}\right) + \log p \cdot \sum_{(i,j):i\neq j} \mathbb{I}\{\sigma(j) < \sigma(i)\} w_{ij}^t = \phi + \log p \cdot c(f_\sigma^{\mathrm{OA}}, \vec{w}^t).
\end{aligned}
$$

Here we define $\phi := \sum_{\ell=1}^t \log\left(\frac{1-p}{1-|p|^{S_\ell}}\right)$, which is independent of σ.

Turning to part 2, let $f, \bar{f} \in \mathcal{M}_p^{OA}$ be such that there exists $\sigma, \bar{\sigma}$ so that $f = f_\sigma^{OA}$ and $\bar{f} = f_{\bar{\sigma}}^{OA}$. As a result,

$$
\begin{aligned}
L_t^{f_\sigma^{\mathrm{OA}}, f_{\bar{\sigma}}^{\mathrm{OA}}} &= \sum_{\ell=1}^{t} \log \frac{f_\sigma^{\mathrm{OA}}(X_\ell | S_\ell)}{\bar{f}_{\bar{\sigma}}^{OA}(X_\ell | S_\ell)} \\
&= \sum_{\ell=1}^{t} \log f_\sigma^{\mathrm{OA}}(X_\ell | S_\ell) - \sum_{\ell=1}^{t} \log f_{\bar{\sigma}}^{\mathrm{OA}}(X_\ell | S_\ell) \\
&= \log p \cdot \left[c(f^{\mathrm{OA}}, \vec{w}^t) - c(f_{\bar{\sigma}}^{\mathrm{OA}}, \vec{w}^t) \right].
\end{aligned}
$$

where the first line follows from part 1 of the proof. This establishes part 2. ■

Proof of Proposition 4. Recall from the proof in Theorem 3 that $\hat{\sigma}_m = (2, \ldots, m, 1, m+1, \ldots, K)$ for $m = 2, \ldots, K$. Moreover, invoking the proof in Theorem 3, we may simplify both max-min problems in Proposition 4 by restricting the alternative preferences to the family of $\{f_{\hat{\sigma}_m}^{\mathrm{OA}} : m = 2, \ldots, K\}$ only. More precisely:

$$
\max_{\lambda \in \Delta(\mathcal{S}^P)} \min_{\bar{f} \in \overline{\mathcal{M}}_p(f_*^{\mathrm{OA}})} D_\lambda\left(f_*^{\mathrm{OA}} || \bar{f}\right) = \max_{\lambda \in \Delta(\mathcal{S}^P)} \min_{m=2,\ldots,K} D_\lambda\left(f_*^{\mathrm{OA}} || f_{\hat{\sigma}_m}^{\mathrm{OA}}\right) = \max_{\lambda \in \Delta(\mathcal{S}^P)} \min_{m=2,\ldots,K} \sum_{S \in \mathcal{S}^P} \lambda(S) d_S\left(\hat{\sigma}_m\right);
$$

$$
\max_{\lambda \in \Delta(\mathcal{S}^{PF})} \min_{\bar{f} \in \overline{\mathcal{M}}_p(f_*^{\mathrm{OA}})} D_\lambda\left(f_*^{\mathrm{OA}} || \bar{f}\right) = \max_{\lambda \in \Delta(\mathcal{S}^{PF})} \min_{m=2,\ldots,K} D_\lambda\left(f_*^{\mathrm{OA}} || f_{\hat{\sigma}_m}^{\mathrm{OA}}\right) = \max_{\lambda \in \Delta(\mathcal{S}^{PF})} \min_{m=2,\ldots,K} \sum_{S \in \mathcal{S}^{PF}} \lambda(S) d_S\left(\hat{\sigma}_m\right).
$$

We will use the shorthand notation $d_S\left(\sigma\right) = D_S\left(f_{\sigma_*}^{\mathrm{OA}} || f_\sigma^{\mathrm{OA}}\right)$ (also used in the proof in Theorem 4) in what follows. Given any randomization $\mu \in \Delta(\{\hat{\sigma}_m : m = 2, \ldots, K\})$, we follow the convention of defining $d_S\left(\mu\right) := \sum_{m=2}^K \mu(\hat{\sigma}_m) d_S\left(\hat{\sigma}_m\right)$. For the remaining of the proof, we show the optimality of $\lambda_*^{\mathrm{P,OA}}$ and $\lambda_*^{\mathrm{PF,OA}}$ separately.

We claim that $\lambda_*^{\mathrm{P,OA}} \in \arg\max_{\lambda \in \Delta(\mathcal{S}^P)} \min_{m=2,\ldots,K} \sum_{S \in \mathcal{S}^P} \lambda(S) d_S\left(\hat{\sigma}_m\right)$. Invoking Lemma 6, for all $i < j \in [K]$,

$$
d_{\{i,j\}}\left(\hat{\sigma}_m\right) = \begin{cases} \frac{\mathfrak{a}_2}{\mathfrak{b}_2} & \text{if } j = m \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \log\left(\frac{1}{p}\right) \frac{1-p}{1+p} & \text{if } j = m \\ 0 & \text{otherwise.} \end{cases}
$$

As a result, for every m, $d_{\lambda_*^{\mathrm{P,OA}}}\left(\hat{\sigma}_m\right) = \frac{1}{K-1} \frac{\mathfrak{a}_2}{\mathfrak{b}_2}$. Consider the randomization $\mu_*^{\mathrm{P,OA}} \in \Delta(\{\hat{\sigma}_m : m = 2, \ldots, K\})$ given by $\mu_*^{\mathrm{P,OA}}(\hat{\sigma}_m) = \frac{1}{K-1}$ for all m. Then $d_S\left(\mu_*^{\mathrm{P,OA}}\right) = \frac{1}{K-1} \frac{\mathfrak{a}_2}{\mathfrak{b}_2}$ for all $S \in \mathcal{S}^P$. Noticing that $d_{\lambda_*^{\mathrm{P,OA}}}\left(\hat{\sigma}_m\right) = d_S\left(\mu_*^{\mathrm{P,OA}}\right)$, we follow the same argument in Theorem 4 and conclude that $\lambda_*^{\mathrm{P,OA}}$ is primal optimal and $\mu_*^{\mathrm{P,OA}}$ is dual optimal for the linear program associated with the max-min problem $\max_{\lambda\in\Delta(\mathcal{S}^P)}\min_{m=2,\ldots,K}\sum_{S\in\mathcal{S}^P}\lambda(S)d_S\left(\hat{\sigma}_m\right)$. (However, unlike Theorem 3, $\lambda_*^{\mathrm{P,OA}}$ is not the unique primal optimal solution.) Correspondingly, the optimal value is

$$\tfrac{1}{K-1}\tfrac{\mathfrak{a}_2}{\mathfrak{b}_2}=(1-p)\log\left(\tfrac{1}{p}\right)\frac{1}{(K-1)(1+p)}.$$

We claim that $\lambda_*^{\mathrm{PF,OA}}\in\arg\max_{\lambda\in\Delta(\mathcal{S}^{PF})}\min_{m=2,\ldots,K}\sum_{S\in\mathcal{S}^P}\lambda(S)d_S\left(\hat{\sigma}_m\right)$. Note that $d_{[K]}\left(\hat{\sigma}_m\right)=\frac{\mathfrak{a}_m}{\mathfrak{b}_K}$ for all $m=2,\ldots,K$. Letting $x=\frac{\mathfrak{a}_2/\mathfrak{b}_2}{\mathfrak{a}_2/\mathfrak{b}_2+(\mathfrak{a}_3-\mathfrak{a}_2)/\mathfrak{b}_K}$, we may verify that

$$d_{\lambda_*^{\mathrm{PF,OA}}}\left(\hat{\sigma}_m\right)=\begin{cases}\frac{\mathfrak{a}_2}{\mathfrak{b}_2}(1-x)+\frac{\mathfrak{a}_2}{\mathfrak{b}_K}x & \text{if } m=2\\ \frac{\mathfrak{a}_m}{\mathfrak{b}_K}x & \text{if } m=3,\ldots,K.\end{cases}$$

Regarding the expression above, we may verify that $\frac{\mathfrak{a}_2}{\mathfrak{b}_2}(1-x)+\frac{\mathfrak{a}_2}{\mathfrak{b}_K}x-\frac{\mathfrak{a}_3}{\mathfrak{b}_K}x=\frac{\mathfrak{a}_2}{\mathfrak{b}_2}-\left(\frac{\mathfrak{a}_2}{\mathfrak{b}_2}-\frac{\mathfrak{a}_3-\mathfrak{a}_2}{\mathfrak{b}_K}\right)x=0$. Plus, since $\mathfrak{a}_n$ strictly increases in n, we may verify that

$$d_{\lambda_*^{\mathrm{PF,OA}}}\left(\hat{\sigma}_K\right)>\cdots>d_{\lambda_*^{\mathrm{PF,OA}}}\left(\hat{\sigma}_4\right)>d_{\lambda_*^{\mathrm{PF,OA}}}\left(\hat{\sigma}_3\right)=d_{\lambda_*^{\mathrm{PF,OA}}}\left(\hat{\sigma}_2\right)=\frac{\mathfrak{a}_3}{\mathfrak{b}_K}x=\frac{\mathfrak{a}_2\mathfrak{a}_3/\mathfrak{b}_2\mathfrak{b}_K}{\mathfrak{a}_2/\mathfrak{b}_2+(\mathfrak{a}_3-\mathfrak{a}_2)/\mathfrak{b}_K}.$$

In the meanwhile, let us consider the randomization $\mu_*^{\mathrm{PF,OA}}\in\Delta(\{\hat{\sigma}_m:m=2,\ldots,K\})$ given by

$$\mu_*^{\mathrm{P,OA}}(\hat{\sigma}_m)=\begin{cases}\frac{\mathfrak{a}_3/\mathfrak{b}_K}{\mathfrak{a}_3/\mathfrak{b}_K+\mathfrak{a}_2(1/\mathfrak{b}_2-1/\mathfrak{b}_K)} & \text{if } m=2\\ \frac{\mathfrak{a}_2(1/\mathfrak{b}_2-1/\mathfrak{b}_K)}{\mathfrak{a}_3/\mathfrak{b}_K+\mathfrak{a}_2(1/\mathfrak{b}_2-1/\mathfrak{b}_K)} & \text{if } m=3\\ 0 & \text{otherwise.}\end{cases}$$

By letting $y=\frac{\mathfrak{a}_3/\mathfrak{b}_K}{\mathfrak{a}_3/\mathfrak{b}_K+\mathfrak{a}_2(1/\mathfrak{b}_2-1/\mathfrak{b}_K)}$ for shorthand notation, we may evaluate $d_S\left(\mu_*^{\mathrm{PF,OA}}\right)$ for every $S\in\mathcal{S}^{PF}$below:

$$d_S\left(\mu_*^{\mathrm{PF,OA}}\right)=\begin{cases}\frac{\mathfrak{a}_2}{\mathfrak{b}_2}y & \text{if } S=[2]\\ \frac{\mathfrak{a}_2}{\mathfrak{b}_K}y+\frac{\mathfrak{a}_3}{\mathfrak{b}_K}(1-y) & \text{if } S=[K].\end{cases}$$

Regarding the expression above, we may verify that $\frac{\mathfrak{a}_2}{\mathfrak{b}_K}y+\frac{\mathfrak{a}_3}{\mathfrak{b}_K}(1-y)-\frac{\mathfrak{a}_2}{\mathfrak{b}_2}y=\frac{\mathfrak{a}_3}{\mathfrak{b}_K}-\left(\frac{\mathfrak{a}_3}{\mathfrak{b}_K}+\frac{\mathfrak{a}_2}{\mathfrak{b}_2}-\frac{\mathfrak{a}_2}{\mathfrak{b}_K}\right)y=0$. That implies $d_{[2]}\left(\mu_*^{\mathrm{PF,OA}}\right)=d_{[K]}\left(\mu_*^{\mathrm{PF,OA}}\right)$. We follow the same argument in Theorem 4 and conclude that $\lambda_*^{\mathrm{PF,OA}}$ is (uniquely) primal optimal and $\mu_*^{\mathrm{PF,OA}}$ is dual optimal for the linear program associated with the max-min problem $\max_{\lambda\in\Delta(\mathcal{S}^{PF})}\min_{m=2,\ldots,K}\sum_{S\in\mathcal{S}^{PF}}\lambda(S)d_S\left(\hat{\sigma}_m\right)$. Correspondingly, the optimal value is

$$\frac{\mathfrak{a}_2\mathfrak{a}_3/\mathfrak{b}_2\mathfrak{b}_K}{\mathfrak{a}_2/\mathfrak{b}_2+(\mathfrak{a}_3-\mathfrak{a}_2)/\mathfrak{b}_K}=(1-p)\log\left(\tfrac{1}{p}\right)\frac{1+2p}{(1-p^K)/(1-p)+2p(1+p)}.$$

This finishes the proof. ■

Proof of Proposition 5. Our main observation in this proof is that we may view the policy F (resp., P and PF) as admissible solutions to our learning problem with an extra constraint about the structure of display sets. With this observation in mind, we will finish this proof by carrying over the arguments in the proofs of Theorems 1, 3, 5, 6, and 7 to the settings of pairwise, full, and pair & full display strategies.

First, let us establish the δ-accuracy of F, P and PF. Note that the proof of Theorem 7 is purely based on a change-of-measure argument restricted to the OA model and then a dominance argument on the "hardness to learn" of the OA model (see also the comments right after Theorem 7). This argument does not rely on the specific structure of the display policy. Hence we may repeat the proof of Theorem 7 verbatim after replacing $\mathcal{S}$ with $\mathcal{S}^F, \mathcal{S}^P$, and $\mathcal{S}^{PF}$ respectively to show that the policies F, P, FP are all δ-accurate.

Second, let us establish the sample complexity of the policies $\{\mathrm{P}, \mathrm{F}, \mathrm{PF}\}$ and conclude that (20) holds. Pick an arbitrary $f' \in \mathcal{M}_p^{\mathrm{OA}}$. In order to show that the estimated preference f_t^{OA} converges to f_*^{OA} fast, let us look at the following three quantities:

- $D_{[K]}\left(f_*^{\mathrm{OA}} || f'\right)$;
- $D_{\lambda_*^{\mathrm{PF,OA}}}\left(f_*^{\mathrm{OA}} || f'\right) \geq \frac{\mathfrak{a}_2/\mathfrak{b}_2}{\mathfrak{a}_2/\mathfrak{b}_2 + (\mathfrak{a}_3 - \mathfrak{a}_2)/\mathfrak{b}_K} \cdot D_{[K]}\left(f_*^{\mathrm{OA}} || f'\right)$;
- $D_{\lambda_*^{\mathrm{P,OA}}}\left(f_*^{\mathrm{OA}} || f'\right) = \frac{\sum_{i=1}^{K-1} \mathbb{I}\{\sigma_{f'}(i) > \sigma_{f'}(i+1)\}}{K-1} \frac{\mathfrak{a}_2}{\mathfrak{b}_2}$.

All of the three quantities above are strictly positive as long as $\sigma_{f'} \neq \sigma_*$. In any case, we know that $f_t^{\mathrm{OA}} \to f_*^{\mathrm{OA}}$ exponentially fast in probability (in the same spirit of the proof of Theorem 6). The rest of the proof is based on repeating the arguments in Theorem 6 by replacing $\mathcal{S}$ with $\mathcal{S}^F$, $\mathcal{S}^P$ and $\mathcal{S}^{PF}$ respectively.

Third, let us establish the lower bound of sample complexity and conclude that (21) holds. We refer to the lower bound result under the general hypothesis testing framework (see Section B.2) and conclude that if we restrict the collection of display sets to $\mathcal{S}^F$ and $\mathcal{S}^P$ respectively, we have that for every $f \in \mathcal{M}_p$,

$$\begin{aligned} \liminf_{\delta \downarrow 0} \frac{\mathbb{E}_f^\pi[\tau]}{\log\left(\frac{1}{\delta}\right)} &\geq \frac{1}{\max_{\lambda \in \Delta(\mathcal{S}^F)} \min_{\bar{f} \in \overline{\mathcal{M}}_p(f)} D_\lambda\left(f || \bar{f}\right)}, \quad \forall \pi \in \mathcal{A}^F; \\ \liminf_{\delta \downarrow 0} \frac{\mathbb{E}_f^\pi[\tau]}{\log\left(\frac{1}{\delta}\right)} &\geq \frac{1}{\max_{\lambda \in \Delta(\mathcal{S}^P)} \min_{\bar{f} \in \overline{\mathcal{M}}_p(f)} D_\lambda\left(f || \bar{f}\right)}, \quad \forall \pi \in \mathcal{A}^P. \end{aligned} \tag{58}$$

Specifically, if $f \in \mathcal{M}_p^{OA}$, and we take π to be F and P respectively, we have $\liminf_{\delta \downarrow 0} \frac{\mathbb{E}_f^F[\tau]}{\log\left(\frac{1}{\delta}\right)} \geq \frac{1}{I^F}$ and $\liminf_{\delta \downarrow 0} \frac{\mathbb{E}_f^P[\tau]}{\log\left(\frac{1}{\delta}\right)} \geq \frac{1}{I^P}$. Align these two inequalities with (20), and we conclude that (21) holds.[16]

[16] Note that a similar result does not apply to the policy PF because the collection of sets $\mathcal{S} = \{[2], [K]\}$ is not closed under permutations. As a consequence, the display sets used by Policy PF may include elements other than [2] and $[K]$, which breaks down the arguments for the lower bound result.

Fourth, let us conclude that the policies {P, F} are worst-case asymptotically optimal for full and pairwise display policies respectively. We notice that the proof of Theorem 3 is based on a dominance argument on the "hardness to learn" of the OA model. Hence we may follow the same argument and conclude that

$$\begin{aligned} I^F &\le \max_{\lambda\in\Delta(\mathcal{S}^F)} \min_{\bar f\in\overline{\mathcal{M}}_p(f)} D_\lambda\left(f||\bar f\right),\ \forall f\in\mathcal{M}_p;\\ I^P &\le \max_{\lambda\in\Delta(\mathcal{S}^P)} \min_{\bar f\in\overline{\mathcal{M}}_p(f)} D_\lambda\left(f||\bar f\right),\ \forall f\in\mathcal{M}_p. \end{aligned}$$

The inequalities above can be taken as equalities when $f\in\mathcal{M}_p^{\mathrm{OA}}$. Hence

$$\begin{aligned} \sup_{f\in\mathcal{M}_p}\liminf_{\delta\downarrow 0}\frac{\mathbb{E}_f^\pi[\tau]}{\log\left(\frac{1}{\delta}\right)} &\ge \sup_{f\in\mathcal{M}_p}\frac{1}{\max_{\lambda\in\Delta(\mathcal{S}^F)}\min_{\bar f\in\overline{\mathcal{M}}_p(f)} D_\lambda\left(f||\bar f\right)} = \frac{1}{I^F}, \quad \forall\pi\in\mathcal{A}^F;\\ \sup_{f\in\mathcal{M}_p}\liminf_{\delta\downarrow 0}\frac{\mathbb{E}_f^\pi[\tau]}{\log\left(\frac{1}{\delta}\right)} &\ge \sup_{f\in\mathcal{M}_p}\frac{1}{\max_{\lambda\in\Delta(\mathcal{S}^P)}\min_{\bar f\in\overline{\mathcal{M}}_p(f)} D_\lambda\left(f||\bar f\right)} = \frac{1}{I^P}, \quad \forall\pi\in\mathcal{A}^P. \end{aligned}$$

In the meantime, (21) implies that

$$\begin{aligned} \sup_{f\in\mathcal{M}_p}\limsup_{\delta\downarrow 0}\frac{\mathbb{E}_f^F[\tau]}{\log\left(\frac{1}{\delta}\right)} &\le \sup_{f\in\mathcal{M}_p}\frac{1}{I^F} = \frac{1}{I^F}\\ \sup_{f\in\mathcal{M}_p}\limsup_{\delta\downarrow 0}\frac{\mathbb{E}_f^P[\tau]}{\log\left(\frac{1}{\delta}\right)} &\le \sup_{f\in\mathcal{M}_p}\frac{1}{I^P} = \frac{1}{I^P}. \end{aligned}$$

Hence the policies {P, F} are worst-case asymptotically optimal for full and pairwise display policies respectively.

Finally, since $\mathcal{S}^F$ is a singleton, $I^F = \max_{\lambda\in\Delta(\mathcal{S}^F)}\min_{\bar f\in\overline{\mathcal{M}}_p(f_*^{\mathrm{OA}})} D_\lambda\left(f_*^{\mathrm{OA}}||\bar f\right) = \min_{\bar f\in\overline{\mathcal{M}}_p(f_*^{\mathrm{OA}})} D_{[K]}\left(f_*^{\mathrm{OA}}||\bar f\right) = \frac{\mathfrak{a}_2}{\mathfrak{b}_K} = (1-p)\log\left(\frac{1}{p}\right)\frac{1}{(1-p^K)/(1-p)}$. The expressions for I^P and I^{PF} can be obtained in the proof of Proposition 4. ■

Proof of Proposition 6. As a consequence of Theorem 6 and Propositions 2 and 5,

$$\lim_{K\uparrow\infty,\ p\uparrow 1}\ \lim_{\delta\downarrow 0}\frac{\mathbb{E}_f^F[\tau]}{\mathbb{E}_f^M[\tau]} \ge \lim_{K\uparrow\infty,\ p\uparrow 1}\frac{K}{K+2p(K-1)}\ \frac{1-p^K}{1-p} = \left(\frac{1}{3}\right)\lim_{K\uparrow\infty,\ p\uparrow 1}\frac{1-p^K}{1-p} \overset{(a)}{=} \left(\frac{1}{3}\right)\lim_{p\uparrow 1}\ \lim_{K\uparrow\infty}\frac{1-p^K}{1-p} = \infty.$$

In the derivations above, equality (a) comes from the fact that mapping $(p,K)\mapsto\frac{1-p^K}{1-p}$ is increasing in both p and K. Similarly, we invoke Theorem 6 and Propositions 2 and 5 again to get

$$\lim_{K\uparrow\infty,\ p\uparrow 1}\ \lim_{\delta\downarrow 0}\frac{\mathbb{E}_f^P[\tau]}{\mathbb{E}_f^M[\tau]} \ge \lim_{K\uparrow\infty, p\uparrow 1}\frac{K}{K+2p(K-1)}(K-1)(1+p) = \infty,$$

which completes the proof. ■

Appendix J: Running Time of the Myopic Tracking Policy

In this section, we provide more implementation details when we evaluate the computational speed of Step 1 of the Myopic Tracking Policy in Section 8.1.

J.1. Two Methods

Integer programming formulation We solve the exact integer programming formulation (15) associated with MTP with the Gurobi MIP solver. Because the simplex method is sometimes significantly slower than the barrier method for the root relaxation of (15), we always use the non-deterministic concurrent algorithm option (provided by Gurobi) to solve the root relaxation.

Our heuristic The heuristic we use combines the idea of majority tournament (with random pivoting rule) in Ailon et al. (2008) and the LP relaxation of (15) as also used in Van Zuylen and Williamson (2009). The LP relaxation of (15) is

$$\begin{aligned} \tilde{x} \in \arg\min_{\vec{x}} & \sum_{(i,j):i\neq j} x_{ji} w_{ij}^t \\ s.t. \quad & x_{ij} + x_{jk} + x_{ki} \geq 1, \quad \forall \text{ distinct } i,j,k \in [K] \\ & x_{ij} + x_{ji} = 1, \quad \forall \text{ distinct } i,j \in [K] \\ & x_{ij} \in [0,1]. \quad \forall \text{ distinct } i,j \in [K] \end{aligned} \tag{59}$$

Given the value of $\tilde{x}$, the estimated ranking is $\tilde{\sigma}^{-1}$, where $\tilde{\sigma}$ is the permutation we obtain using Algorithm 3.

Algorithm 3 LP-Rand-Pivot

```
  INPUT: Collection of Items $\mathcal{C} \subseteq [L]$
 STEP 0: Randomly pick a pivot $k \in V$
 STEP 1: $\mathfrak{L} = \{i \in [K] \setminus \{k\} : \tilde{x}_{ki} \leq 0.5\}$, $\mathfrak{R} = \{i \in [K] \setminus \{k\} : \tilde{x}_{ki} > 0.5\}$
OUTPUT: LP-Rand-Pivot($\mathfrak{L}$), $k$, LP-Rand-Pivot($\mathfrak{R}$)
        (Concatenation of left recursion, $k$, and right recursion.)
```

To get some intuition about Algorithm 3, note that a large value of $\tilde{x}_{ij}$ represents a stronger tendency that item i is ranked higher than (or preferred to) item j. Hence $\mathfrak{L}$ (resp. $\mathfrak{R}$) in Algorithm 3 corresponds to items that are ranked higher (resp. lower) than the pivot item k.

Let us also say a few words on how our heuristic is connected to the earlier literature. The difference between our heuristic and that in Ailon et al. (2008) is that the underlying tournament in our heuristic is $A_H = \{(i,j) : \tilde{x}_{ij} > \tilde{x}_{ji}\}$, while the underlying tournament in Ailon et al. (2008) is $A_M = \{(i,j) : w_{ij}^t > w_{ji}^t\}$. The main difference between our heuristic and that in Van Zuylen and Williamson (2009) (among others) is that we do not solve an underlying optimization problem to find the pivot item k. The reason we implement this heuristic is that this heuristic combines the simple implementation of Ailon et al. (2008) (e.g., random pivoting, majority tournament) and the information of the solution to the LP relaxation. In particular, if the LP returns an integral solution, our heuristic is guaranteed to achieve zero optimality gap.

J.2. Simulation Details

We generate representative instances of (15) that arise as we implement MTP. Our simulation is based on two parts. In the first part, we iterate over the learning problem instances by taking $p \in \{0.5, 0.9\}$ and K ranging from 25 to 150. For each problem instance (p, K), we use our heuristic (for Step 1) and the optimal display randomization (for Step 3) to generate a 500 period run of MTP five different times (allowing for random realizations of display sets and consumer choices) and track the instances of (15) that arise along the way. The underlying choice model is the OA model. We record the computation times and optimality gaps and aggregate them at the level of (K, p). Here the optimality gap is defined as $(v_H - v_{LP})/v_H$, where v_H is the objective function value of our heuristic and v_{LP} is the optimal value of the LP relaxation of (15). Note that the optimality gap relative to the LP relaxation is always an upper bound of the "actual" gap compared to the exact solution.

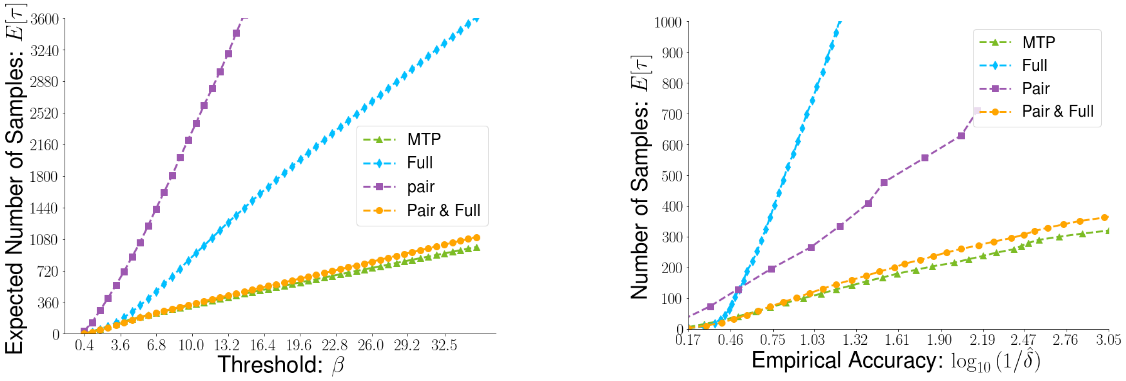
In the second part, we evaluate the computational performance of exactly solving the integer programming formulation. We do so by revisiting the IP instances generated in the first part with two different approaches: (i) we solve those (same) instance exactly using the Gurobi MIP solver (instead of approximately using our heuristic); and (ii) we focus on a subset of periods where $t \in \{20, 40, 60, \ldots, 500\}$ so as to speed up the simulation process. We record the computation times and aggregate them at the level of (K, p).

Appendix K: AGH Survey Data

In this appendix, we test the performance of our proposed MTP policy using a real data set from a student survey conducted at AGH University of Science and Technology. Specifically, we have sampled from the (first) AGH Course survey dataset available from PREFLIB, an online library of datasets concerning preferences (see PREFLIB 2019). This is a dataset with the complete rankings of 146 students regarding eight course modules, collected at AGH University of Science and Technology, Krakow, in 2003.

To emulate our sequential learning setting, we randomly generate students from the data set and present them with a subset of courses. We use a rather intuitive approach to convert the available complete ranking data to a choice model (see, e.g., Désir et al. 2018). Specifically, given any display set, we first (uniformly) randomly draw a complete ranking from the dataset (with replacement), and then pick the course that is ranked highest in the display set. This conversion rule gives rise to a particular preference model f_{AGH}, where $f_{AGH}(X|S)$ represents the empirical fraction of students who rank course X at the top within display set S. We run the same four policies consider in Section 8 under the preference model f_{AGH} to see which uses the least amount of samples (i.e., choices) to recover the students' most preferred course with high probability. It is important to note that f_{AGH} does not satisfy Assumption (A-3) in Definition 1, i.e, the notion of "ground truth ranking" is not well defined. It does, however, satisfy the weaker condition (A-5) discussed in Section 9 and so lies in $\widetilde{\mathcal{M}}_p$. Indeed, we do find that there is a course i_* that is chosen with higher probability than any other courses under f_{AGH} regardless of the display set. In other words, the problem of identifying the "ground truth most preferred course" is well defined.

In Figure 9, we look at the sample complexity of the four policies as a function of (a) the stopping threshold β and (b) the resulting empirical error probability (or the fraction of instances where the algorithm terminates with a course different from i_*). In the implementation of these experiments, we adopt the integer programming formulation of the MLE problem and set $p = 0.7$.



(a) stopping threshold β

(b) empirical error probability

Figure 9 Comparison of policy performance on the AGH Survey Data

These numerical results suggest that the Myopic Tracking Policy dominates the three benchmark policies, stopping earlier for every fixed threshold β and for every value of the empirical error probability. This suggests that the Myopic Tracking policy is effective, even in settings more general than the one analyzed in the paper.