Ranking Items from Discrete Ratings: The Cost of Unknown User Thresholds

Oscar Villemaud OSCAR.VILLEMAUD@EPFL.CH

Suryanarayana Sankagiri SURYANARAYANA.SANKAGIRI@EPFL.CH

Matthias Grossglauser MATTHIAS.GROSSGLAUSER@EPFL.CH

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

Ranking items is a central task in many information retrieval and recommender systems. User input for the ranking task often comes in the form of ratings on a coarse discrete scale. We ask whether it is possible to recover a fine-grained item ranking from such coarse-grained ratings. We model items as having scores and users as having thresholds; a user likes an item if the score exceeds the threshold, and dislikes it otherwise. Although all users implicitly agree on the total item order, estimating that order is challenging when both the scores and the thresholds are latent. Under our model, any ranking method naturally partitions the n items into bins; the bins are ordered, but the items inside each bin are still unordered. Users arrive sequentially, and every new user can be queried to refine the current ranking. We prove that achieving a near-perfect ranking, measured by Spearman distance, requires $\Theta(n^2)$ users (and therefore $\Omega(n^2)$ queries). This is significantly worse than the $O(n \log n)$ queries needed to rank either from comparisons or from ratings with known user thresholds; the gap reflects the additional queries needed to estimate each user's latent threshold. Our bound also quantifies the impact of a mismatch between the score and threshold distributions via a quadratic divergence factor. To show the tightness of our results, we provide a ranking algorithm whose query complexity matches our bound up to a logarithmic factor. Our work reveals a tension in online ranking: diversity in thresholds is necessary to merge coarse ratings from many users into a fine-grained ranking, but this diversity has a cost if the thresholds are a priori unknown.

Keywords: Ranking, Lower Bound, Discrete Ratings, Threshold Model

1. Introduction

Ranking items according to human preferences is a central task underlying many digital platforms: search engines order links, recommender systems curate content, and peer review selects papers for publication. In many such applications, the feedback signal is a discrete score for each item: a binary action like a click or a like, or ratings on a small integer scale. In this paper, we ask the question: how hard is it to obtain fine-grained rankings from coarse ratings?

Surprisingly, this fundamental question has received scant attention in the theoretical machine learning literature. Most prior work treats discretization as a noisy signal of the item's score (*i.e.*, utility). In contrast, we view discretization as the result of a thresholding process. That is, we assume that each user has a *latent discretization threshold*, and they rate two items differently only when this threshold lies between the items' scores. Thus, a platform can order two items only by querying a user who has an appropriately placed threshold. In most circumstances, this discretization threshold is a priori unknown to the platform, which makes it difficult to order and rank items.

In contrast, asking users to compare items eliminates the effect of discretization: any user can tell the order between two items. In many practical learning-to-rank tasks, it has indeed been argued that asking users to compare items may be preferable to soliciting ratings, if possible.

In this work, we study the aforementioned phenomena by developing a simple probabilistic model of a platform that aims to rank items from user-provided ratings. In most of the analysis, we assume binary ratings, but we comment on the extension to finer discretizations later. We assume that there are n items, with each item i having a score X_i (or utility) in the interval $[0, 1]$. Users arrive sequentially to the system, and the platform adaptively chooses which items to query to the user. Each user u has a latent discretization threshold $Y_u \in [0, 1]$. When asked to rate an item i, user u rates it as 1 if and only if the item's score exceeds the user's threshold ($X_i > Y_u$). We assume the item scores and user thresholds to be i.i.d. in $[0, 1]$, of respective densities f_X and f_Y. For simplicity, we assume users do not return to the system once they have finished answering the platform's rating queries.

This model assumes that all users agree on the order of the items, even though they may have different *rating styles*: users with a low threshold are more lenient, while users with a higher threshold are more stringent. The model makes it clear why multiple users are necessary; if we ask the first user to rate all items, it still splits them only into two bins; every successive user can at best split one existing bin into two new bins, and so on. In this way, at any given stage, the method maintains a *partial ranking* over the items, successively refining it with every new user.

This suggests that in order to rank all items, we need a diverse population of users with different discretization thresholds. In particular, the population of thresholds, as it were, should be dense in the same region where the items are closely spaced. Furthermore, the information added by every new user is subject to diminishing returns: at best, a new user partitions a single current bin, whose size decreases over time. The information added per sample (query) decreases even more quickly: as the task progresses, it becomes increasingly costly to locate the single bin that the new user can partition.

A basic question is: how many users must the platform see in order to recover the underlying ranking? In our model, this number is a stopping time. Somewhat surprisingly, we find that the expected number of users needed for a perfect ranking is infinite, *irrespective of the number of items* (see Lemma 1). Consequently, we instead quantify the accuracy of the partial ranking resulting from a fixed number of users. We measure the accuracy in terms of the Spearman footrule distance, a natural metric to compare two rankings.

Our main results, Theorems 5 and 6, identify the sample complexity of ranking in two different regimes. Theorem 5 shows that if the number of users m scales linearly with the number of items n, then the estimated ranking differs from the ground-truth by a footrule distance of order $O(n)$. Loosely, this means that each item's rank can be identified to within a constant distance. Theorem 6 extends this analysis and shows that if the number of users m scales super-linearly with n, then the average footrule distance scales as $O(n^2/m)$. In particular, if $m = \Omega(n^2)$, then the platform can rank all but a few items correctly. In other words, to obtain a near-perfect ranking of n items, one needs $\Omega(n^2)$ users.

Our analysis also quantifies to what extent a misalignment between the score and threshold distributions increases the sample complexity; in both theorems, the dominant term contains a quadratic divergence factor $\mathbb{E}[(f_X(Y)/f_Y(Y))^2]$, which is smallest when f_X and f_Y are identical, and increases as the two distributions drift apart. This is intuitive: user thresholds add the most information if they tend to fall where there are items to be partitioned, and are wasted otherwise.

These results should be interpreted as lower bounds on the complexity of ranking from discrete ratings. The number of users is clearly a lower bound for the sample (*i.e.*, query) complexity; indeed, a user who is never queried at least once adds neither cost nor any information. This sample complexity result should be contrasted with the complexity of ranking from pairwise comparisons. It is well-known that sorting algorithms can obtain a complete ranking with $O(n \log n)$ (adaptively-chosen) pairwise comparisons. Our model shows that ranking items by soliciting user ratings under latent thresholds is fundamentally much harder than ranking them by asking users to compare items. This handicap is made even worse if the score and threshold distributions are poorly matched. Note that the issue is indeed that the user thresholds are unknown: if every new threshold Y_u were available to the system for free, then methods can be devised whose sample complexity is close to that of efficient sorting algorithms. The difference stems from the queries that need to be expended for every user to, in effect, estimate its threshold Y_u (or equivalently, to find the bin it can partition).

In addition to the theoretical analysis outlined above, we provide insights into the tightness of our lower bounds by designing and analyzing an algorithm, called threshold binary search (TBS), for latent-threshold ranking. The algorithm uses a binary search–style strategy to steer each user's queries toward the most informative regions. We give a proof sketch of the algorithm's complexity, and find that it matches the lower bounds up to logarithmic factors. These findings are also corroborated through simulation results.

In conclusion, our work provides the first clean theoretical explanation for a widely observed empirical fact: comparisons are more effective than ratings for ranking. These results have practical implications. If fine-grained rankings are needed, it is more efficient to solicit comparisons rather than ratings. If the context requires ratings, systems should ensure diversity across users' thresholds and design queries so that users are most discriminative in regions where items are concentrated. We discuss extensions to k-ary ratings and noisy ratings in the discussion section of the paper.

Structure of the paper Section 2 formalizes our model and presents the problem. Section 3 proves a relation between the number of users and the expected precision of the ranking. In Section 4, we study the tightness of our results by providing an algorithm that efficiently orders items from binary ratings. Section 5 shows experimental results that support the conclusions of the previous two sections. Finally, we discuss related work and conclude in Section 6.

2. Model and Problem

We consider a model with a finite number n of items. Each item $i \in [n]$ has an unknown score X_i that represents its utility. The item scores X_i are *iid* random variables on $[0, 1]$, of density f_X. There is a potentially infinite number of users who give feedback on the items by rating them. Although the item scores are common to all users, the rating of different users for the same item may differ because different users have different *rating styles*. This rating style manifests itself in the form of a unique discretization threshold for ratings. Thus, some users rate most items as 1 while others rate most items as 0. The rating style of a user $u \in \mathbb{N}$ is represented by an individual threshold $Y_u \in [0, 1]$ that controls their rating in the following way:

$$\forall u \in \mathbb{N}, \forall i \in [n], q(u, i) \triangleq \mathbb{1}(X_i > Y_u)$$

where $q(u, i)$ is the binary rating given by user u when queried with item i . Under this model, the item score is the probability that a random user rates the item as 1. The thresholds Y_u are *iid* uniform random variables in $[0, 1]$, of density f_Y. We assume that f_X and f_Y are non-zero, upper bounded and c-Lipschitz.

2.1. The Cost Of Exact Ranking

We assume that users arrive in a sequence and are asked to rate some selected items according to an algorithm. Given a set of items, our goal is to study how many users and queries we need to order the items. Consider a fixed set of users E. Under our model, it is possible to relatively order two items i and j if and only if there exists a user $u \in E$ such that $X_i < Y_u < X_j$ (or $X_i > Y_u > X_j$). This means that a necessary condition to fully order the items is to have a user threshold between all pairs of consecutive item scores. If the items were equally spaced in the $[0, 1]$ interval, the number of thresholds would correspond to the classical coupon collector problem. Then, the expected number of users would be $\mathbb{E}[M] = n \log(n)$. However, the items scores are *iid* on $[0, 1]$, so the number of users needed to order all items is infinite in expectation, as stated by the following lemma:

Lemma 1 *Let $X_1, \ldots, X_n$ be iid item scores of density f_X on $[0, 1]$. Let M be the random number of users needed to obtain a total order. Then we have:*

$$\mathbb{E}[M] = \infty$$

Proof [Idea of proof for $f_X = f_Y = 1$.] Consider an interval between two consecutive items. Conditional on this interval, the expected number of user thresholds we need to try until one falls in this interval and separates the two items is inversely proportional to its length. The problem arises because the distribution of the interval length is a Beta of parameter one, *i.e.*, does not vanish at zero. In other words, we are too likely to get at least one very short interval, which dominates the total cost. We formally prove this result in Appendix C. ■

In summary, Lemma 1 suggests that a more interesting regime is one where we assume a finite number of users, and ask how well we can approximate the ground-truth ranking. In order to formalize the notion of partial order for our problem, we introduce a new structure (the bin sequence) and an associated error metric (the MSF).

2.2. The Problem of Partial Ranking

As stated before, we need a threshold between each pair of consecutive item scores in order to fully order the items, and this requires an infinite expected number of users. We now assume that we have a finite number of users, and we observe what we can say about the order of the items. Consider the first user and assume they rate all items. This gives us the information about which items have their score below Y_1 and which have their score above. This divides the items in two groups such that we know the relative ordering of a pair of items (i, j) *if and only if* i and j are not in the same group. If i and j are in the same group, we have no information on the order of their scores. This is the maximum amount of information the first user can provide. With more users, we are able to divide the items in more groups, but items within the same group are still indistinguishable. In the rest of the paper, we refer to these groups of items as *bins*, and to the ordered set of bins as a *bin sequence*.

Definition 2 ***Bins and Bins Sequences.*** *We call* ***bin sequence*** *an ordered partition $\mathfrak{B} = (\mathcal{B}_1, ..., \mathcal{B}_{|\mathfrak{B}|})$ of $[n]$ that respects the order of the item scores. Formally:*

$$\forall k < k' \in [|\mathfrak{B}|], \forall i \in \mathcal{B}_k, i' \in \mathcal{B}_{k'}, X_i < X_{i'}$$

$\mathcal{B}_1, \ldots, \mathcal{B}_{|\mathfrak{B}|}$ are called ***bins*** *of items. They are sets of items whose scores belong to a same interval. We denote the partial order induced by this bin sequence by $\mathcal{S}_{\mathfrak{B}}$,* i.e. *$\mathcal{S}_{\mathfrak{B}}$ is the set of orderings compatible with $\mathfrak{B}$.*

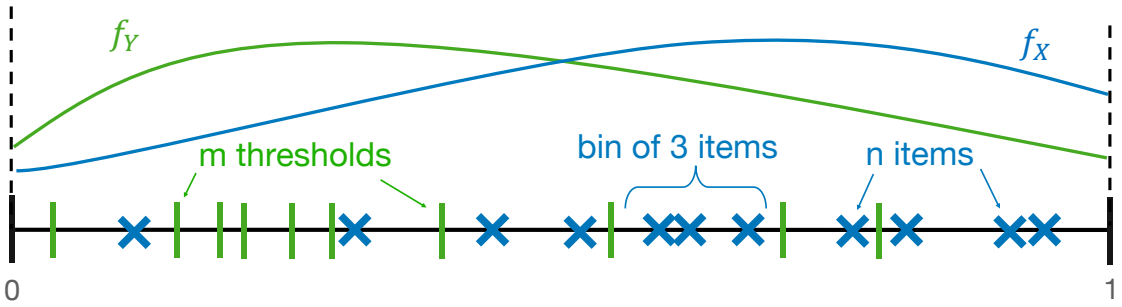


Figure 1: Item scores and user thresholds are sampled *iid*, respectively with densities f_X and f_Y.

Our model and the notion of bin are illustrated in Figure 1. In the following, when we talk about partial order, we always refer to a partial order induced by a bin sequence. In order to define the precision of a partial order, we rely on the *Spearman Footrule* (SF) distance between two permutations to define the *Maximum Spearman Footrule* (MSF). For any finite subset $\mathcal{B}$ of $\mathbb{N}$, let $\mathcal{S}_\mathcal{B}$ denote the set of possible orders of $\mathcal{B}$. For any order $\sigma \in \mathcal{S}_\mathcal{B}$ and for any $i \in \mathcal{B}$, $\sigma(i)$ denotes the rank of item i among the elements of $\mathcal{B}$. Then the SF distance between two orders on $\mathcal{B}$ is defined as:

Definition 3 ***Spearman Footrule.*** $\forall \sigma, \sigma^* \in \mathcal{S}_\mathcal{B}$, $SF(\sigma, \sigma^*) \triangleq \sum_{i \in \mathcal{B}} |\sigma(i) - \sigma^*(i)|$

Given a partial order, the MSF measures the worst (*i.e.* maximum) Spearman Footrule between two total orders compatible with the partial order:

Definition 4 ***Maximum Spearman Footrule (MSF).***
Let $\mathcal{B} \subseteq \mathbb{N}$ be a finite set. Let $\mathcal{S}$ be a subset of $\mathcal{S}_\mathcal{B}$. Then,

$$MSF(\mathcal{S}) \triangleq \max_{\sigma, \sigma' \in \mathcal{S}} SF(\sigma, \sigma')$$

In simple terms, the MSF is the "diameter" of the set of possible orderings, with respect to the Spearman Footrule distance. In particular, if the true ordering σ^* belongs to $\mathcal{S}$, then we have $\forall \sigma \in \mathcal{S}, \quad \mathtt{SF}(\sigma, \sigma^*) \leq \mathtt{MSF}(\mathcal{S})$.

Because the bin sequence and the MSF are fully determined by the set of item scores and user thresholds, the MSF is a random variable which is independent of the choice of an algorithm (assuming the algorithm extracts all the information from the users). In what follows, we denote this random variable by F.

In the rest of the paper, we study how many users and items are needed to obtain a constant expected MSF. A constant MSF means that most items are correctly placed, and only a constant number of them has uncertainty on their rank. Allowing for this vanishing fraction of items not to be fully ordered is what allows us to overcome the impossibility result of Lemma 1. Additionally, we study an easier version of the problem, where we allow for a linear MSF in the number of items. This second setting corresponds roughly to having all items within a fixed distance of their true rank.

3. Expected Maximum Spearman Footrule

In this section we compute $\mathbb{E}[F]$, the expected MSF, given the number of users n and the number of items m, where the expectation is taken with respect to the randomness of the item scores and the user thresholds. Theorems 5 and 6, which are our main results, respectively give the expected MSF when the number of users and items are of the same order and when we have asymptotically more users than items. We give ideas of proof in Section 3.2 and full proofs in Appendix D.

3.1. Main Theorems

Recall that n is the number of items, m is the number of users, and f_X and f_Y are their densities. Let Y be a threshold selected uniformly at random (so Y also has density f_Y).

Theorem 5 *Assume that there exists $r \in \mathbb{R}^+$ s.t. $m \sim rn$ as n goes to infinity, then*

$$\left|\mathbb{E}[F] - n\left(\frac{1}{2} + \frac{1}{r}\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right]\right)\right| \leq \frac{r}{2}n + o(n)$$

Theorem 6

Assume that there exists $r \in \mathbb{R}^+, \gamma > 1$ s.t. $m \sim rn^\gamma$ as n goes to infinity, then

$$\mathbb{E}[F] \sim \frac{2}{r}n^{2-\gamma}\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right]$$

Interpretation of the theorems Theorem 5 shows that, if the number of items grows linearly with the number of users, then the expected MSF also grows linearly. Theorem 6 shows in particular that we need to take $\gamma = 2$, *i.e.* a quadratic number of users, if we want to keep a constant MSF. If the user thresholds were known, an efficient algorithm could discard users that do not provide additional information without asking any rating from them. But, in our case, it is pointless to reject a user who has not provided any rating, because all users who have not rated anything are indistinguishable. Consequently, the number of users constitutes a natural lower bound of the number of queries.

Influence of the distributions The divergence term $\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right]$ appears in both theorems. This expression captures the effect of having different distributions for the items and the thresholds. Indeed, the MSF scales quadratically with the size of the bins (see Lemma 9). This means that, for a given number of items and users, the smallest MSF is achieved when the thresholds are equally spaced between the items. Additionally, we can guess that having a high concentration of items in a region without thresholds will have a strong effect on the MSF, whereas the converse is not true. This asymmetry appears in the mathematical expression, because we have $\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right] = \int_0^1 \frac{f_X(y)^2}{f_Y(y)}dy$, which shows that this term will blow up if for instance the support of f_Y is smaller than the one of f_X. In Appendix G, Lemma 27, we show by convexity that this divergence is greater than 1, the value 1 being reached for $f_X = f_Y$. In addition, we show in Lemma 28 that in the case where scores follow a $Beta(a_X, b_X)$, and thresholds a $Beta(a_Y, b_Y)$, we have the closed form $\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right] = \frac{\mathbf{B}(a_Y,b_Y)}{\mathbf{B}(a_X,b_X)^2}\mathbf{B}(2a_X - a_Y, 2b_X - b_Y)$, where $\mathbf{B}$ is the Beta function.

3.2. Proof of Theorems

We present the main results used for Theorems 5 and 6. Full proofs are given in Appendix D and E.

Our model allows for general densities for both item scores and user thresholds. However, we can observe that composing all scores and thresholds by an increasing function on $[0,1]$ would leave the order of the items and thresholds unchanged. This shows that there exists different pairs (f_X, f_Y) that have the same properties with respect to our problem. In particular, from any (f_X, f_Y), we can compose everything by the *cdf* F_Y, in order to have a uniform distribution of the thresholds on $[0,1]$. Thus, any (f_X, f_Y) model has an equivalent $(f_{X'}, 1)$ model, with the same properties when it comes to the MSF. Therefore, in the rest of the section, we assume that $f_Y = 1$, *i.e.* the user thresholds are uniform *iid* in $[0,1]$. Under this assumption, our results are expressed in function of $\mathbb{E}[f_X(Y)^2]$. We can show that this term becomes the $\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right]$ of Theorems 1 and 2 for $f_Y \neq 1$. We formally explain this manipulation in Appendix G.1 and prove the full result in Lemma 26.

We now present our main lemmas, proven under the assumption $f_Y = 1$. The results of Theorems 5 and 6 both come from the following lemma, which splits the expected MSF in two terms.

Lemma 7 *Let B be the number of items in a bin chosen uniformly at random among all the bins. Then,*

$$\mathbb{E}[F] = \frac{1}{2}(m+1)(\mathbb{E}[B^2] - \mathbb{P}(B \text{ is odd}))$$

This result comes from the following two lemmas (8 and 9, proven in Appendix D.1). The first shows that the whole MSF can be decomposed in the sum of the MSF of each bin, and the second expresses the MSF of a bin as a function of the number of items in the bin. Recall that if $\mathfrak{B}$ is a bin sequence and $\mathcal{B}$ is a bin, $\mathcal{S}_{\mathfrak{B}}$ is the partial order associated with $\mathfrak{B}$, and $\mathcal{S}_{\mathcal{B}}$ is the set of all possible orderings of $\mathcal{B}$. For instance, if $\mathfrak{B} = (\mathcal{B}_1)$ (trivial bin sequence with only one bin), then $\mathcal{S}_{\mathfrak{B}} = \mathcal{S}_{\mathcal{B}_1}$.

Lemma 8 *The MSF of a partial order is the sum of the MSF of each bin.*

$$\mathtt{MSF}(\mathcal{S}_{\mathfrak{B}}) = \sum_{k=0}^{m} \mathtt{MSF}(\mathcal{S}_{\mathcal{B}_k})$$

Lemma 9 *Let $\mathcal{B}$ be a bin (*i.e. *a finite set).*
Then $\mathtt{MSF}(\mathcal{S}_{\mathcal{B}}) = \frac{|\mathcal{B}|^2}{2}$ if $|\mathcal{B}|$ is even and $\mathtt{MSF}(\mathcal{S}_{\mathcal{B}}) = \frac{|\mathcal{B}|^2-1}{2}$ if $|\mathcal{B}|$ is odd.

Lemma 7 shows that the MSF can be computed from $\mathbb{E}[B^2]$ and $\mathbb{P}(B \text{ is odd})$, where B is the size of a bin selected uniformly at random. These two terms are computed in the following two lemmas:

Lemma 10 *Let $\beta \in (0.5, 1)$. Then,*

$$\mathbb{E}[B^2] = \frac{n}{m+1} + 2\frac{n^2-n}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{m^{2\beta+1}}\right)$$

Lemma 11 *Let $\beta \in (0.5, 1)$, $m = \omega(n)$. Then, for n going to infinity,*

$$\mathbb{P}(B \text{ is odd}) = \frac{n}{m} - 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + O\left(\frac{n^3}{m^3}\right)$$

Lemmas 10 and 11 are results on the number of items in a random bin. In order to prove them, we need to study the distribution of the number of items in a bin, which is hard to derive directly. However, its distribution conditioned on the two extremities of the bin is easier to derive. For all $k \in [m]$, let $D_k \triangleq Y_{k+1} - Y_k$ be the *length* of the bin (with $Y_0 = 0$ and $Y_{m+1} = 1$). We further define $\forall k \in [m], \quad P_k \triangleq \mathbb{P}(X_i \in [Y_k, Y_{k+1}] | Y_k, D_k) = \int_{Y_k}^{Y_k + D_k} f_X(x)dx$. The expression does not depend on i, because the item scores are *iid*, and independent of the thresholds. Furthermore, the number of items in bin k conditioned on (Y_k, D_k) follows a binomial distribution of parameters (n, P_k). This is because P_k is the conditional probability that an item is in bin k. In order to compute $\mathbb{E}[B^2]$ in Lemma 10, we need the values of $\mathbb{E}[P_k^2]$ for all k. We use Lemma 12 to approximate these values.

Lemma 12 *$\forall \beta \in (0.5, 1)$, when m goes to infinity,*

$$\mathbb{E}\left[\sum_{k=0}^{m} P_k^2\right] = \sum_{k=0}^{m} \mathbb{E}[(D_k f_X(Y_k))^2] + O\left(\frac{1}{m^{2\beta}}\right)$$

In order to prove Lemma 12, we define a probabilistic event $\mathcal{E}(\beta, m) \triangleq \left(\bigcap_{k=0}^{m} \left(D_k \leq \frac{1}{m^\beta}\right)\right)$ (see Appendix F), whose probability goes exponentially to 1 as m goes to infinity (Lemma 24). Then, we show that $\mathcal{E}(\beta, m)$ implies that P_k is well approximated by $D_k f_X(Y_k)$ (Lemma 25). The remaining terms $\mathbb{E}[(D_k f_X(Y_k))^2]$ are computed in Appendix E.

4. Upper Bound On The Number of Queries

In this section we provide some intuition on the tightness of our lower bound. For this, we present an algorithm that solves the task of partial ranking from ratings, and analyze its complexity in the case $f_X = f_Y$. We provide experiments on the empirical complexity of our algorithm in Section 5.

As explained in Section 2, our model assumes that users arrive sequentially to the system, in an arbitrary order. Thus, for each user, nothing is known initially about their threshold, beyond their prior distribution. For each user, the algorithm adaptively chooses a sequence of items to query. The algorithm can choose the items based on its current state, which depends on all the responses it has gathered till then (including the current user's past responses). The algorithm may choose any number of items to query the user with, and observes the corresponding noiseless ratings.

The algorithm we propose maintains a bin sequence representing all the information about the order of the items, which is updated with every new user. For each user u, it extracts all the information they can provide. This means finding out which item scores are smaller than Y_u and which are bigger. In order to do so, we make the user rate all items of the bin that contains Y_u. The position of the other item scores relative to Y_u can be inferred because the bins are ordered in the bin sequence.

In order to find the bin containing Y_u bin with the least number, we rely on the following observation: if the new user gives a rating of 1 to an item of bin $\mathcal{B}_k$, it implies that they would also give a rating of 1 to all items of any bin $\mathcal{B}_l$, with $l > k$. So the index of the bin containing Y_u has to be smaller or equal to k. This shows that it is possible to perform a binary search on the bins, with a small subtlety: for each bin selected during the binary search, we only learn if the index of the bin containing the threshold is *smaller or equal* or *greater or equal*. Consequently, using only one query per bin, it is possible to isolate a pair of consecutive bins out of which one is the correct one, but it is not possible to know which of the two.

Using this idea, we present our algorithm TBS (Threshold Binary Search), which is detailed in Algorithm 2. The execution of TBS is divided in steps, each step corresponding to a different user. Each step is divided in three phases : SEARCH, ISOLATE and SPLIT. These phases are detailed in Algorithms 3, 4 and 5 in Appendix H.1. In what follows, we refer by $\texttt{SEARCH}_u$, $\texttt{ISOLATE}_u$, $\texttt{SPLIT}_u$ and $\texttt{UPDATE}_u$ to the execution of these phases at step u.
Let Y_u be the (unknown) threshold of user u. $\texttt{SEARCH}_u$ corresponds to the binary search described before. It finds a subset of two adjacent bins of which one of the two contains Y_u (or Y_u is between the two bins). $\texttt{ISOLATE}_u$ identifies which of these two bins is the one actually containing Y_u by alternating the queries between the two bins (if Y_u is between the bins, the biggest one is returned). $\texttt{SPLIT}_u$ splits the selected bin in two new bins by requesting the user to rate all items of the bin. Finally, we update the bin sequence by replacing the bin being split by the two new bins. If one of the two bins returned by $\texttt{SPLIT}_u$ is empty, it means that the new user does not bring additional information. In this case, the bin sequence is not updated.

Upper bound on the complexity of TBS We do not formally prove the complexity of our algorithm, but for the case where the densities of the scores and thresholds are the same and by making some reasonable asymptotic approximations, we are able to show that the expected number of queries needed is $O(n\log(m) + m\log(n))$.

Idea of proof We split the total number of queries Q in $Q = Q^{search} + Q^{split} + Q^{iso}$, the cost of each of the three phases of the algorithm. For each user, the SEARCH phase cannot make more than $\log_2(n)$ queries, because it is a binary search on the sequence of nonempty bins. So we have $Q^{search} \leq m\log_2(n)$. By using the fact that ISOLATE defaults to the biggest of the two bins, we can show that $Q^{iso} \leq Q^{split}$. Estimating the complexity of Q^{split} is the most difficult part of the analysis. To do it, for each user-item pair (i, u), we look at the probability that the u-th user treated by the algorithm rates i during the SPLIT phase. We show that this probability is roughly $\frac{1}{u}$, which gives a total cost of $\mathbb{E}[Q^{split}] \simeq \sum_{i=1}^{n}\sum_{u=1}^{m}\frac{1}{u} \simeq n\log(m)$. Combining these three results gives us the upper bound $\mathbb{E}[Q] = O(n\log(m) + m\log(n))$. We detail this argument in Appendix H.2.

Interpretation In our cases of interest, where we have at least as many users as items, the dominating term is $m\log(n)$, *i.e.* a logarithmic number of queries per user. So, in these regimes, our lower bound (which relies on the number of users m) is optimal up to a logarithmic factor.

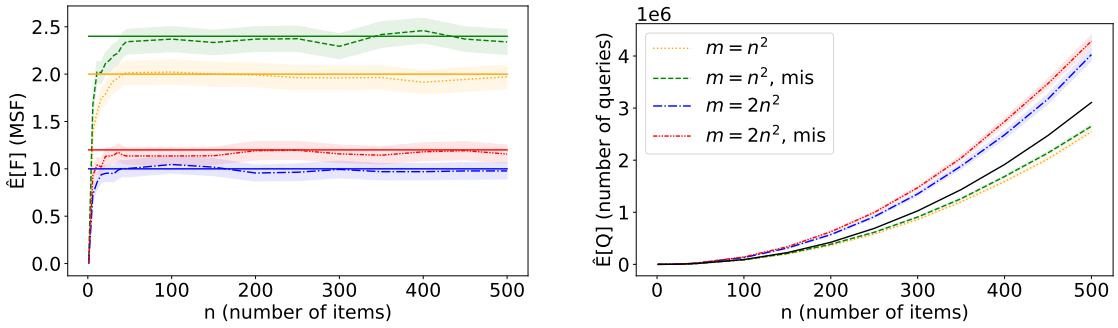
Algorithm 1: $\texttt{TBS}(n, m)$

$\mathcal{B}_1 \leftarrow [n]$; /* bin */
$\mathfrak{B} \leftarrow (\mathcal{B}_1)$; /* bin sequence */
for u *going from* 1 *to* m **do**
  $l, r \leftarrow \texttt{SEARCH}(\mathfrak{B}, u)$; /* find pair of bins containing Y_u (Algo 3) */
  $k^* \leftarrow \texttt{ISOLATE}(\mathfrak{B}, l, r, u)$; /* identify the correct one (Algo 4) */
  $\mathcal{B}^-, \mathcal{B}^+ \leftarrow \texttt{SPLIT}(\mathcal{B}_{k^*}, u)$; /* try to split the bin in two (Algo 5) */
  if $|\mathcal{B}^-| > 0$ *and* $|\mathcal{B}^+| > 0$ **then**
    $\mathfrak{B} \leftarrow (\mathcal{B}_1, ..., \mathcal{B}_{k^*-1}, \mathcal{B}^-, \mathcal{B}^+, \mathcal{B}_{k^*+1}, ..., \mathcal{B}_{|\mathfrak{B}|})$
  end
end
Return $\mathfrak{B}$

5. Experiments

We perform experiments[1] to support our claim of the previous section and show that the convergence of Theorem 6 is fast enough. Experiments related to Theorem 5 are available in Appendix I.

Experiment setting We run Monte-Carlo experiments using Python. All experiments are run 1000 times and we report the average results on the plots along with the 95% confidence intervals. We fix the randomness using seeds 1 to 1000 for reproducibility. All experiments were ran using an Apple M3 Pro chip with 18GB of RAM. Our experiments are constructed as follows. First, we generate our data by sampling n *iid* item scores from a $Beta(a_X, b_X)$ and m *iid* user thresholds following a $Beta(a_Y, b_Y)$. Second, we run `TBS` on this data until the algorithm terminates. Finally, we plot our metrics (MSF obtained and number of queries made). We display the average taken over the different runs, along with the 95% confidence interval (*e.g.* $\hat{\mathbb{E}}[F] \pm 1.96\frac{\sqrt{\hat{\mathbb{V}}(F)}}{\sqrt{s}}$, where s is the number of samples, and $\hat{\mathbb{E}}$ and $\hat{\mathbb{V}}$ are the empirical mean and variance). The MSF is computed by taking the bin sequence returned by the algorithm and using Lemmas 8 and 9 of Section 3.2.



(*a*) Expected MSF

(*b*) Expected number of queries

Figure 2: Experiments on the expected MSF and number of queries for a number of users quadratic in the number of items. The full lines on Figure (a) are the theoretical values of Theorem 6. Figure (b) shows the empirical query cost of `TBS`. The solid black line shows $y = 2n^2\log(n)$, which corresponds to the rate estimated in Section 4 (the constant 2 is manually chosen to roughly match the other lines). The label *'mis'* indicates the experiments with a mismatch between the distributions of the items and the thresholds. We use $a_X = 2, b_X = 3, a_Y = 2, b_Y = 2$ for the mismatch case, and $a_X = 1, b_X = 1, a_Y = 1, b_Y = 1$ in the default case.

Interpretation of results We see on Figure 2(*a*) that the empirical MSF reaches the limit predicted by Theorem 6 at around 50 items. This shows that our asymptotic results can be used to estimate the expected MSF even for a small number of items. We observe on Figure 2(*b*) that the number of queries of `TBS` follows the $m\log(n)$ (*i.e.* $n^2\log(n)$ here) tendency discussed in Section 4. Interestingly, we can also see that the mismatch increases the average number of queries per user, but only by a small margin. These results support our theoretical analysis of Sections 3 and 4 by providing strong evidence that our bounds are tight up to a logarithmic factor.

1. Our code is available at https://anonymous.4open.science/r/threshold_model_alt_26

6. Discussion and Related Work

This work addresses the problem of adaptive ranking from discrete ratings. Prior work on adaptive ranking has primarily relied on pairwise comparisons to order items (Jamieson and Nowak, 2011; Heckel et al., 2019), with some studies highlighting the efficiency gains of settling for approximate rankings (Heckel et al., 2018). Classical sorting algorithms such as QuickSort also rely on comparisons. A wide range of comparison oracle models have been proposed, but none interpret comparisons as being derived from discrete ratings. Our model introduces a new kind of oracle: a comparison is only available when a user's threshold separates the scores of the two items. We show that ranking items under this oracle is significantly harder than even a noisy comparison oracle like Plackett-Luce.

A closely related work is by Garg and Johari (2019), which also tackles the problem of inferring a ranking from discrete ratings. Like we do for our algorithm, they emphasize the importance of posing the "right question". However, their model differs significantly: they assume a homogeneous user population that provides noisy ratings based on item utilities, and their focus is on selecting the best query to pose to the entire population. In contrast, we consider a heterogeneous population and focus on identifying the right user whose threshold yields informative ratings. Medo and Wakeling (2010) also explore the effects of discretization in recommender systems. Although their model shares structural elements with ours, *e.g.*, items with latent scores on a unit interval and ratings derived via thresholding—they assume uniform thresholds and focus on algorithmic consequences within collaborative filtering. In contrast, we explore how unknown, user-specific thresholds complicate the problem of ranking items even when users agree on the ground-truth ordering.

More broadly, our work is motivated by a long-standing interest in how best to elicit feedback in recommender and crowdsourcing systems. This question has appeared under different names, including *type of feedback* (Babski-Reeves et al., 2023), *method of elicitation* (Shah et al., 2016), and *format* (Fernandes et al., 2023). A central debate concerns the trade-offs between cardinal (rating-based) and ordinal (comparison-based) feedback. Ratings are easier to aggregate across users and are the default in many systems. However, growing empirical evidence supports the superiority of comparisons in several respects: they tend to be less biased (Fernandes et al., 2023), more consistent over time (Jones et al., 2011), and cognitively less demanding (Shah et al., 2016; Xu et al., 2024). Studies have shown that user rating biases are pervasive and prone to drift over time (Harik et al., 2009), complicating aggregation. Although some works (Wang and Shah, 2019) suggest that cardinal feedback can outperform ordinal feedback, these rely on continuous rating scales—a setting fundamentally different from ours. Our results complement these findings, and should be interpreted in the light of works like the ones of Sparling and Sen (2011) and Jones et al. (2011), who study the time taken, cognitive load or user satisfaction when using different types of feedbacks.

By isolating the cost of discretization in ratings, our work provides new insights into the debate between of ordinal and cardinal feedback. Indeed, we show that even in the absence of noise, the cost incurred by the discretization is leading to a high complexity for ratings, higher than the one of comparisons. We showed that even if we had a priori knowledge that a group of users agree on the underlying item ranking, extracting that ranking efficiently is costly because we need to expend queries on every new user to learn about her unknown threshold, *before* we can extract useful information from that user to contribute to the estimated ranking. Although we made a number of simplifying assumptions in this model, we believe that the broader observation on rating feedback is

valuable: the fact that users have diverse thresholds necessitates spending a large number of queries to determine whether they can indeed order items.

Main takeways The most important implication of our work is that when fine-grained ranking is required, *comparison-based feedback* is significantly more sample-efficient than *discrete ratings*. As we show, $O(n \log n)$ rating queries suffice to estimate ranks within a constant deviation per item (linear MSF), but reducing this deviation further requires a much larger number of queries — due to diminishing marginal information gain from each new user. Our work provides one interpretation of the commonly held belief: soliciting comparisons instead of ratings is better for ranking items. A second practical takeaway of this work is that one suffers a penalty when the distribution of item scores is mismatched with the distribution of user thresholds. In this work, we have quantified this penalty via the term $\mathbb{E}[(f_X(Y)/f_Y(Y))^2]$, which grows larger as the two distributions become more different. In practical systems, this suggests that *knowing the score distribution* and designing feedback prompts to make users most discriminative around that region is crucial. This insight complements the work of Garg and Johari (2019), who also study the design of optimal binary prompts for ranking (e.g., asking "Is this essay better than average?" versus "Is it better than good?").

Limitations Our model assumes that items have a fixed utility shared across users and that the only source of randomness is the distribution of discretization thresholds. Relaxing either assumption, for example, by incorporating noise in ratings or allowing for user disagreements, would likely make the ranking problem harder. As such, our lower bounds should still hold in broader settings, albeit possibly in weaker forms. The restriction to binary ratings is not fundamental. Our lower bounds generalize to k-level rating scales. Indeed, if we make the assumption that each of the m users has $k-1$ *iid* thresholds in $[0,1]$ instead of only one threshold, the distribution of the set of all thresholds is the same as having $m(k-1)$ users with one threshold each. In this case, the lower bound on the number of users needed scales down by a factor of $k-1$. Finally, our algorithm is tightly coupled to our model assumptions and may not perform well under real-world noise or preference heterogeneity. Its primary role is to match our lower bounds and establish tightness. That said, the algorithm's core idea—using binary search to isolate sets of items for a user to rate—offers a practical design principle that may generalize to more robust algorithms in future work.

References

K Babski-Reeves, B Eksioglu, and D Hampton. Inferring user preferences using cardinal vs. ordinal feedback in recommender systems. *IISE Annual Conference.Proceedings*, 2023.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1643–1668, 2023.

Nikhil Garg and Ramesh Johari. Designing optimal binary rating systems. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1930–1939. PMLR, 2019.

Polina Harik, Brian E Clauser, Irina Grabovsky, Ronald J Nungester, Dave Swanson, and Ratna Nandakumar. An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, 46(1):43–58, 2009.

Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin Wainwright. Approximate ranking from pairwise comparisons. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1057–1066. PMLR, 2018.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, and Martin J Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 2019.

Kevin G Jamieson and Robert Nowak. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.

Nicolas Jones, Armelle Brun, and Anne Boyer. Improving reliability of user preferences: Comparing instead of rating. In *2011 Sixth International Conference on Digital Information Management*, pages 316–321, September 2011.

Matúš Medo and Joseph Rushton Wakeling. The effect of discrete vs. continuous-valued ratings on reputation and ranking systems. *Europhysics Letters*, 91(4):48004, 2010.

Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramch, Martin J Wainwright, et al. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

E Isaac Sparling and Shilad Sen. Rating: how difficult is it? In *Proceedings of the fifth ACM conference on Recommender systems*, pages 149–156, 2011.

Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 864–872, 2019.

Austin Xu, Andrew McRae, Jingyan Wang, Mark Davenport, and Ashwin Pananjady. Perceptual adjustment queries and an inverted measurement paradigm for low-rank metric learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix A. Outline

We provide here some intuitions and pointers to help the reader navigate through the Appendix.

A.1. On the expected MSF

Scaling of the MSF Our main results, Theorems 5 and 6, are proven in Appendix D. In essence, they show that the MSF scales as $\frac{n}{m^2}$, where n is the number of items and m the number of users. Both theorems rely on Lemma 7, which shows that the MSF grows quadratically with the bin sizes (number of items in each bin), and linearly with the number of bins (*i.e.* the number of users plus one). Lemma 7 also shows that the parity of the bin sizes has a small bounded effect on the MSF. This effect is a consequence of the dichotomy of Lemma 9. In practice, this means that the parity of the bin size has a significant effect only when the MSF is not going to infinity. For this reason, deriving the probability that bins contain an odd number of items is not needed for Theorem 5, which deals with high MSF regimes. On the contrary, Theorem 6 deals with the cases where the MSF is small. Therefore, its proof requires additional analysis to take into account the probability of bins sizes being odd.

Properties of the bins The number of items in the k-th bin B_k depends on both the length of the bin D_k and the density of the items f_X on the $[Y_k, Y_{k+1}]$ interval. As the number of thresholds m goes to infinity, we can use the Lipschitzness of f_X to argue that f_X is roughly constant of value $f_X(Y_k)$ over the bin. This idea is formalized in Appendix F. Because they are the differences of the order statistics of *iid* independent random variables, the bin lengths $(D_k)_k$ follow Beta distributions (Lemma 22). Although the $(D_k)_k$ are not independent, they essentially converge to *iid* exponential random variables as m goes to infinity. However, for finite m analysis, we need to take into account not only the dependency between the $(D_k)_k$, but also the dependency between the $(D_k)_k$ and the $(Y_k)_k$. For this reason, we work with conditional distributions. In particular, we prove several results on the length D of a bin selected uniformly at random conditional on the left extremity Y of the bin (Lemmas 19, 20, 21).

Border effects Although the results would be the same by overlooking it, rigorous analysis has to treat differently the first bin, which has a deterministic threshold on its left side ($Y_0 = 0$). Indeed, when considering the set of all bins $\{B_k | k \in [m] \cup \{0\}\}$ and their left extremities $\{Y_k | k \in [m] \cup \{0\}\}$, the first threshold Y_0 is not an order statistic of *iid* random variables. This means that this special case must be treated separately, which makes an additional term appear in the proofs, although it turns out to be negligible. Because of this subtlety, Y (a threshold selected uniformly at random) does not exactly have density f_Y, but rather converges in law to density f_Y as m goes to infinity. Understanding this can help the reader process the proofs with more ease.

A.2. On the complexity of the `TBS` algorithm

In Appendix H, we study the complexity of our algorithm. We recall here the main ideas, already presented in Section 4. The analysis relies on splitting the total number of queries Q in $Q = Q^{search} + Q^{split} + Q^{iso}$, the cost of each of the three phases of the algorithm. For each user, the `SEARCH` phase cannot make more than $\log_2(n)$ queries, because it is a binary search on the sequence of nonempty bins. So we have $Q^{search} \leq m \log_2(n)$. By using the fact that `ISOLATE` defaults to the biggest of the two bins, we can show that $Q^{iso} \leq Q^{split}$. Estimating the complexity of

Q^{split} is the most difficult part of the analysis. To do it, for each user-item pair (i, u), we look at the probability that the u-th user treated by the algorithm rates i during the SPLIT phase. We show that this probability is roughly $\frac{1}{u}$, which gives a total cost of $\mathbb{E}[Q^{split}] \simeq \sum_{i=1}^{n}\sum_{u=1}^{m}\frac{1}{u} \simeq n\log(m)$. Combining these three results gives us the upper bound $\mathbb{E}[Q] = O(n\log(m) + m\log(n))$.

A.3. Structure of the Appendix

In Appendix B, we recall our definitions and introduce new notation useful to our analysis. In Appendix D, we prove our main results. In particular we prove some preliminary lemmas in Appendix D.1 before proving Theorem 5 in Appendix D.2 and Theorem 6 in Appendix D.3. In Appendix E, we present various properties of the random variables of our model, which are used to prove Theorems 5 and 6. In Appendix F, we define a probabilistic event under which the lengths of all bins are uniformly upper bounded. In Appendix G, we show why we can assume the thresholds follow a uniform distribution without loss of generality, and we present some results on the quadratic divergence. In Appendix H, we present the detailed idea of proof of the complexity of TBS. In Appendix I, we present additional experiments. In Appendix J, we provide a few lemmas on inequalities used in the appendix.

Appendix B. Definitions and Notation

B.1. Main Notation

We summarize here the detailed notation of the random variables used in our theoretical results. Throughout the paper, we use the following notation:

- n: number of items
- m: number of users
- f_X: the density of the item scores on $[0, 1]$. Scores are *iid* on the interval.
- $f_Y = 1$: the density of the user thresholds on $[0, 1]$. Thresholds are *iid* on the interval.
- $\forall i \in [n], X_i$ is the (unordered) score of item i. X_is are *iid* of density f_X.
- $\forall k \in [m], Y_k$ is the k-th smallest threshold, *i.e.* the k-th order statistic of m *iid* random variables of density f_Y (with convention $Y_0 = 0$, and $Y_{k+1} = 1$).
- $\forall k \in [m] \cup \{0\}, D_k \triangleq Y_{k+1} - Y_k$ the length of bin number k.
- $\forall k \in [m] \cup \{0\}, B_k \triangleq |\{i \in [n] | X_i \in [Y_k, Y_{k+1}]\}|$, the number of items in bin k.
- $B \triangleq B_K, K \sim \mathcal{U}([m] \cup \{0\})$, the number of items in a random bin, chosen uniformly at random among the bins. We define in the same way $Y \triangleq Y_K$ and $D \triangleq D_K$, a random threshold and the length of a random bin.
- F: the (random) value of the MSF.

Note that the $f_Y = 1$ assumption is without loss of generality, as detailed in Appendix G.1.

B.2. Definition of the MSF

We recall here the definitions necessary to define the MSF. First, we define the need the following preliminary notation:

- For any finite subset $\mathcal{B}$ of $\mathbb{N}$, $\mathcal{S}_\mathcal{B}$ is the set of possible orders of $\mathcal{B}$.
- For any order $\sigma \in \mathcal{S}_\mathcal{B}$ and for any $i \in \mathcal{B}$, $\sigma(i)$ is the rank of item i among the elements of $\mathcal{B}$.

Using this notation, we can define bins, bin sequences, the Spearman footrule and the MSF.

Definition 2 ***Bin Sequence and Bins.*** *We call* ***bin sequence*** *an ordered partition $\mathfrak{B} = (\mathcal{B}_1, ..., \mathcal{B}_{|\mathfrak{B}|})$ of $[n]$ that respects the order of the item scores.*
Formally :

$$\forall k < k' \in [|\mathfrak{B}|], \forall i \in \mathcal{B}_k, i' \in \mathcal{B}_{k'}, X_i < X_{i'}$$

$\mathcal{B}_1, \ldots, \mathcal{B}_{|\mathfrak{B}|}$ are called ***bins*** *of items. They are sets of items whose scores belong to a certain interval.*
We note $\mathcal{S}_\mathfrak{B}$ the partial order induced by this bin sequence, i.e. *$\mathcal{S}_\mathfrak{B}$ is the set of orderings compatible with $\mathfrak{B}$.*

Definition 3 ***Spearman Footrule.*** $\forall \sigma, \sigma^* \in \mathcal{S}_\mathcal{B}$, $SF(\sigma, \sigma^*) \triangleq \sum_{i \in \mathcal{B}} |\sigma(i) - \sigma^*(i)|$

Given a partial order, the MSF measures the worst (*i.e.* maximum) Spearman Footrule between two total orders compatible with the partial order:

Definition 4 ***Maximum Spearman Footrule (MSF).***
Let $\mathcal{B} \subseteq \mathbb{N}$ be a finite set. Let $\mathcal{S}$ be a subset of $\mathcal{S}_\mathcal{B}$. Then,

$$MSF(\mathcal{S}) \triangleq \max_{\sigma, \sigma' \in \mathcal{S}} SF(\sigma, \sigma')$$

B.3. Other Definitions

We define here more advanced notation which is used in the proofs.

Definition 13 (P_k, P, $p(Y, D)$) *We define P_k as the conditional probability that an item is in bin k:*

$$\forall i \in [n], \forall k \in [m] \cup \{0\}, \quad P_k \triangleq \mathbb{P}(X_i \in [Y_k, Y_{k+1}] | Y_k, D_k) = \int_{Y_k}^{Y_k + D_k} f_X(x)dx \triangleq p(Y_k, D_k)$$

In the same way as for the other random variables depending on k, we define $P \triangleq P_K$, where K is uniform on $[m] \cup \{0\}$.

Remark The expression does not depend on i, because the item scores are *iid* and independent of the thresholds.

We provide here the definition of event $\mathcal{E}$, which is discussed more in detail in Appendix F.

Definition 14 ($\mathcal{E}(\beta, m)$) *For all $\beta \in (0, 1), m \in \mathbb{N}$, we define the event $\mathcal{E}(\beta, m)$ as follows:*

$$\mathcal{E}(\beta, m) \triangleq \left(\bigcap_{k=0}^{m} \left(D_k \leq \frac{1}{m^\beta} \right) \right)$$

Appendix C. On the Difficulty of Perfect Ranking

Lemma 1 *Let $X_1, \dots, X_n$ be iid item scores following a $\mathcal{U}([0,1])$ distribution. Let M be the random number of users needed to obtain a total order. Then we have*

$$\mathbb{E}[M] = \infty$$

Proof As discussed in Section 3.2, and proved in Appendix G.1, we can make the assumption that either $f_X = 1$ or $f_Y = 1$ without loss of generality. Here, we assume $f_X = 1$, *i.e.* the item scores are uniformly distributed on $[0,1]$.
For all $i \in [n]$, let us note $X_{(i)}$ the i-th smallest score. Let $U_{1,2}$ be the random variable of the number of users sampled before the arrival of a threshold in $[X_{(1)}, X_{(2)}]$. Clearly, $U_{1,2}$ is smaller than the total number of users sampled before having a threshold between every pair of consecutive items. So we have $\mathbb{E}[M] \geq \mathbb{E}[U_{1,2}]$. Let us show $\mathbb{E}[U_{1,2}] = \infty$. Let us define $c_Y \triangleq \max f_Y$. We have

$$\begin{aligned}
\mathbb{E}[U_{1,2}] &= \mathbb{E}\left[\mathbb{E}[U_{1,2} \mid X_{(1)}, X_{(2)}]\right] \\
&= \mathbb{E}\left[\frac{1}{\int_{X_{(1)}}^{X_{(2)}} f_Y(y)dy}\right] \\
&\geq \mathbb{E}\left[\frac{1}{\int_{X_{(1)}}^{X_{(2)}} c_Y dy}\right] \\
&= \frac{1}{c_Y}\mathbb{E}\left[\frac{1}{X_{(2)} - X_{(1)}}\right] \\
&= \frac{1}{c_Y}\int_0^1 \frac{1}{x} n(1-x)^{n-1} dx \\
&= +\infty
\end{aligned}$$

The second equality comes from the fact that the thresholds are independent in $[0,1]$, so the number of users needed to obtain a threshold in $[X_{(1)}, X_{(2)}]$ follows a geometric random variable of parameter $p = \int_{X_{(1)}}^{X_{(2)}} f_Y(y)dy$.

The last two equalities are because $X_{(2)} - X_{(1)}$ is the difference of consecutive order statistics of a uniform distribution, which follows a Beta distribution $Beta(1, n)$ (Lemma 22), and finally the integral diverges at 0.

■

Discussion on Lemma 1. The intuition is that M, the number of users needed until we get a perfect ordering, is a random variable whose distribution has a very heavy tail: it has a decay of $1/m^2$. Equivalently, the complementary CDF (i.e. $\mathbb{P}(M > m)$), has a decay of $1/m$. It is easy to show that such a distribution has infinite mean; it follows from the fact that the series $\sum_m 1/m$ is not summable.

Digging a little deeper sheds more light on why the CDF of has such a heavy tail. The fundamental result is that we need a user threshold between every two item scores in order to rank the

items fully. In the simple case when there are only two items, we need a user threshold between them. Assume for the sake of this argument that both item scores and user thresholds are uniformly distributed. Let the two item scores be laid out first. Given that the gap between the two item scores is , the probability that a random user's threshold lies between these two is x. For each user, this event is independent. Therefore, the expected number of users to pass by until this random event happens is $1/x$. Now, the gap between two items could be any number between zero and one. The exact distribution of the gap is a beta random variable, but the important point is that this distribution has strictly positive density at zero. Therefore, the probability of the gap being between x and $x + dx$ is proportional to dx. Integrating $c(1/x)dx$ from zero to one gives us infinity. A slightly misleading line of argument is the following. When both the user thresholds and item scores are uniformly distributed, by symmetry, there is a $1/3$ chance that a user threshold lies between the item scores. Since these events are independent, on average, it should take 3 users to land a user whose threshold lies between the score. The fallacy in this argument is that the events: "the threshold of user u lies between the two item scores" are not independent across users. The fact that the first user's threshold did not lie between the scores gives us some information that the item scores are probably closer together than what is known a priori. This reduces the chance that the second user's threshold lies between the scores. We get independence of these events only by conditioning on the item scores. Arguably, this simple counter-intuitive fact is what makes this problem so interesting to study.

Appendix D. Main Results

In this section, we prove our two main results, Theorems 5 and 6 in Appendix D.2 and D.3, which give the expected MSF in function of the number of items and users. Before this, we prove the main lemmas needed to prove our theorems in Appendix D.1.

D.1. Preliminary Results to the Theorems

In this section, we show that the expected MSF depends on two terms: $\mathbb{E}[B^2]$ and $\mathbb{P}(B \text{ is odd})$. Then, we compute those two terms, thus dealing with most of the difficulties of Theorems 5 and 6.

D.1.1. EXPRESSING THE MSF IN FUNCTION OF B

Lemma 7

$$\mathbb{E}[F] = \frac{1}{2}(m+1)(\mathbb{E}[B^2] - \mathbb{P}(B \textit{ is odd}))$$

Proof

$$\begin{aligned}
\mathbb{E}[F] &= \frac{1}{2}\sum_{k=0}^{m}(\mathbb{P}(B_k \text{ is even})\mathbb{E}[B_k^2 \mid B_k \text{ is even}] + \mathbb{P}(B_k \text{ is odd})(\mathbb{E}[B_k^2 - 1 \mid B_k \text{ is odd}])) \\
&= \frac{1}{2}\sum_{k=0}^{m}\left(\mathbb{E}[B_k^2] - \mathbb{P}(B_k \text{ is odd})\right) \\
&= \frac{1}{2}(m+1)\left(\mathbb{E}[B^2] - \mathbb{P}(B \text{ is odd})\right)
\end{aligned}$$

where the first equality uses Lemmas 8 and 9.

■

Lemma 7 is proven using the following two lemmas, which are deterministic results on the MSF in the context of bins of items.

Lemma 8 *The MSF of a partial order is the sum of the MSF of each bin,* i.e.*, for $\mathfrak{B}$ a bin sequence,*

$$MSF(\mathcal{S}_{\mathfrak{B}}) = \sum_{k=1}^{|\mathfrak{B}|} MSF(\mathcal{S}_{\mathcal{B}_k})$$

Proof Let us define the MSF of a set with respect to an ordering $\sigma^* \in \mathcal{S}$:

$$\texttt{MSF}(\mathcal{S}, \sigma^*) \triangleq \max_{\sigma \in \mathcal{S}} \texttt{SF}(\sigma, \sigma^*)$$

For all $\sigma \in \mathcal{S}_{\mathfrak{B}}$, let $\sigma_{|\mathcal{B}}$ be the order induced by σ on bin $\mathcal{B}$. Let us prove the following preliminary result:

$$\forall \sigma^* \in \mathcal{S}_{\mathfrak{B}}, \texttt{MSF}(\mathcal{S}_{\mathfrak{B}}, \sigma^*) = \sum_{k=1}^{|\mathfrak{B}|} \texttt{MSF}(\mathcal{S}_{\mathcal{B}_k}, \sigma^*_{|\mathcal{B}_k}) \tag{1}$$

Let $\mathfrak{B} = (\mathcal{B}_1, \ldots, \mathcal{B}_{|\mathfrak{B}|})$ be the bin sequence. Let $\sigma^* \in \mathcal{S}_{\mathfrak{B}}$. Let $\sigma^- = \arg\max_{\sigma \in \mathcal{S}_{\mathfrak{B}}} SF(\sigma, \sigma^*)$ be the order for which the MSF is reached. Then σ^- is the worst possible ordering that respects this bin sequence. We have:

$$\begin{aligned}
\texttt{MSF}(\mathcal{S}_{\mathfrak{B}}, \sigma^*) &= SF(\sigma^-, \sigma^*) \\
&= \sum_{i \in [n]} |\sigma^*(i) - \sigma^-(i)| \\
&= \sum_{\mathcal{B} \in \mathfrak{B}} \sum_{i \in \mathcal{B}} |\sigma^*(i) - \sigma^-(i)|
\end{aligned}$$

For any $\sigma \in \mathcal{S}_{\mathfrak{B}}$, $\forall k \in [|\mathfrak{B}|], \forall i \in \mathcal{B}_k$, we have:

$$\sigma_{|\mathcal{B}_k}(i) = \sigma(i) - g(k(i))$$

where, $g(k(i)) = \sum_{l<k(i)} |\mathcal{B}_l|$ is the number of items in all bins before the one containing item i.

This gives us :

$$\forall \mathcal{B} \in \mathfrak{B}, \forall i \in [mi], |\sigma^*_{|\mathcal{B}}(i) - \sigma^-_{|\mathcal{B}}(i)| = |(\sigma^*(i) - g(k(i))) - (\sigma^-(i) - g(k(i)))| = |\sigma^*(i) - \sigma^-(i)|$$

so $\forall \mathcal{B} \in \mathfrak{B}$,

$$\begin{aligned}
\sum_{i \in \mathcal{B}} |\sigma^*(i) - \sigma^-(i)| &= \sum_{i \in \mathcal{B}} |\sigma^*_{|\mathcal{B}}(i) - \sigma^-_{|\mathcal{B}}(i)| \\
&= \texttt{MSF}(\mathcal{S}_{\mathcal{B}}, \sigma^*_{|\mathcal{B}})
\end{aligned}$$

So $\texttt{MSF}(\mathcal{S}_{\mathfrak{B}}, \sigma^*) = \sum_{\mathcal{B}\in\mathfrak{B}} \texttt{MSF}(\mathcal{S}_{\mathcal{B}}, \sigma^*_{|\mathcal{B}})$, thus proving equation (1).
Using this results, we have:

$$\begin{aligned}
\texttt{MSF}(\mathcal{S}_{\mathfrak{B}}) &\triangleq \max_{\sigma,\sigma^*\in\mathcal{S}_{\mathfrak{B}}} \texttt{SF}(\sigma,\sigma^*) && (2)\\
&= \max_{\sigma^*\in\mathcal{S}_{\mathfrak{B}}} \max_{\sigma\in\mathcal{S}_{\mathfrak{B}}} \texttt{SF}(\sigma,\sigma*) && (3)\\
&= \max_{\sigma^*\in\mathcal{S}_{\mathfrak{B}}} \texttt{MSF}(\mathcal{S}_{\mathfrak{B}},\sigma^*) && (4)\\
&= \max_{\sigma^*\in\mathcal{S}_{\mathfrak{B}}} \sum_{k=1}^{|\mathfrak{B}|} \texttt{MSF}(\mathcal{S}_{\mathcal{B}_k}, \sigma^*_{|\mathcal{B}_k}) && (5)\\
&= \sum_{k=1}^{|\mathfrak{B}|} \max_{\sigma^*\in\mathcal{S}_{\mathcal{B}_k}} \texttt{MSF}(\mathcal{S}_{\mathcal{B}_k},\sigma^*) && (6)\\
&= \sum_{k=1}^{|\mathfrak{B}|} \max_{\sigma^*\in\mathcal{S}_{\mathcal{B}_k}} \max_{\sigma\in\mathcal{S}_{\mathcal{B}_k}} \texttt{SF}(\sigma,\sigma^*) && (7)\\
&= \sum_{k=1}^{|\mathfrak{B}|} \max_{\sigma,\sigma^*\in\mathcal{S}_{\mathfrak{B}_k}} \texttt{SF}(\sigma,\sigma^*) && (8)\\
&= \sum_{k=1}^{|\mathfrak{B}|} \texttt{MSF}(\mathcal{S}_{\mathcal{B}_k}) && (9)
\end{aligned}$$

where the equality of line (6) is true because any order σ on the full set can be decomposed in a sequence of one order on each bin. ■

Lemma 9 *Let $\mathcal{B}$ be a bin (*i.e. *a finite set).*
Then $\texttt{MSF}(\mathcal{S}_{\mathcal{B}}) = \frac{|\mathcal{B}|^2}{2}$ if $|\mathcal{B}|$ is even and $\texttt{MSF}(\mathcal{S}_{\mathcal{B}}) = \frac{|\mathcal{B}|^2-1}{2}$ if $|\mathcal{B}|$ is odd.

Proof For $b \in \mathbb{N}$, we write MSF_b the value of $\texttt{MSF}(\mathcal{S}_{\mathcal{B}})$ for $b = |\mathcal{B}|$.
Let us first consider the case where $\mathcal{B}$ is even.
The value of MSF_b is equal to the Spearman footrule between two opposite orderings of $\mathcal{B}$. The displacement between the first and the last element is $b-1$. The displacement between the second and the second to last elements is $b-3$. This goes on until the displacement of 1 on the two central elements. Consequently:

$$MSF_b = (b-1) + (b-3) + \ldots + 3 + 1 + 1 + 3 + \ldots + (b-3) + (b-1) = \frac{b^2}{2}$$

Using this, we can prove the result for b odd. If the number of elements is even, and they are in reverse order (such that the SF is maximal), then adding an new element in the middle creates a new reverse ordering, and it adds one to the SF of each of the previous elements (the new element has an SF of 0 because it is in the middle). So, $\forall k \in \mathbb{N}$

$$MSF_{2k+1} = MSF_{2k} + 2k = \frac{4k^2+4k}{2} = \frac{(2k+1)^2-1}{2}$$

■

D.1.2. COMPUTING $\mathbb{E}[B^2]$

In this section, we prove Lemma 10 and its corollary.

Lemma 10 *Let $\beta \in (0.5, 1)$. Then*

$$\mathbb{E}[B^2] = \frac{n}{m+1} + 2\frac{n^2-n}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{m^{2\beta+1}}\right)$$

Proof

We use Lemmas 12, 18 and 21, all proven Appendix E. We have

$$\sum_{k=0}^{m}\mathbb{E}[B_k^2] = n + (n^2-n)\sum_{k=0}^{m}\mathbb{E}[P_k^2] \tag{10}$$

$$= n + (n^2-n)\left(\sum_{k=0}^{m}\mathbb{E}[D_k^2 f_X(Y_k)^2] + O\left(\frac{1}{m^{2\beta}}\right)\right) \tag{11}$$

$$= n + (n^2-n)\left((m+1)\mathbb{E}[D^2 f_X(Y)^2] + O\left(\frac{1}{m^{2\beta}}\right)\right) \tag{12}$$

where we used Lemma 18 for the first equality and Lemma 12 for the second. Furthermore, we have

$$\mathbb{E}[D^2 f_X(Y)^2] \tag{13}$$

$$= \mathbb{E}[f_X(Y)^2\mathbb{E}[D^2|Y]] \tag{14}$$

$$= \mathbb{E}[f_X(Y)^2\left(\mathbb{E}[D^2|Y, Y\neq 0]\mathbb{P}(Y\neq 0) + \mathbb{E}[D^2|Y, Y=0]\mathbb{P}(Y=0)\right)] \tag{15}$$

$$= \mathbb{E}\left[f_X(Y)^2\left(2\frac{1-Y^{m+1}-(m+1)Y^m(1-Y)}{(m+1)m}\frac{m}{m+1} + \frac{2}{(m+1)(m+2)}\frac{1}{m+1}\right)\right] \tag{16}$$

$$= \frac{2}{(m+1)^2}\mathbb{E}\left[f_X(Y)^2\left(1-Y^{m+1}-(m+1)Y^m(1-Y)\right)\right] + \frac{2}{(m+1)^2(m+2)} \tag{17}$$

where (16) comes from Lemma 21.

In addition, using $c_X \triangleq \max f_X$, we have

$$\begin{aligned}
\mathbb{E}[f_X(Y)^2] &\geq \mathbb{E}\left[f_X(Y)^2\left(1 - Y^m - mY^{m-1}(1-Y)\right)\right] \\
&= \mathbb{E}[f_X(Y)^2] - \mathbb{E}[f_X(Y)^2 Y^m] - m\mathbb{E}[f_X(Y)^2 Y^{m-1}(1-Y)] \\
&\geq \mathbb{E}[f_X(Y)^2] - c_X^2\mathbb{E}[Y^m] - c_X^2 m\mathbb{E}[Y^{m-1}(1-Y)] \\
&= \mathbb{E}[f_X(Y)^2] - c_X^2\frac{1}{m+1} - c_X^2 m\frac{1}{m(m+1)} \\
&= \mathbb{E}[f_X(Y)^2] - \frac{2c_X^2}{m+1}
\end{aligned}$$

so, combining with (17),

$$\frac{2}{(m+1)^2}\left(\mathbb{E}[f_X(Y)^2] - \frac{2c_X^2}{m+2}\right) + \frac{2}{(m+1)^2(m+2)} \leq \mathbb{E}[D^2 f_X(Y)^2] \leq \frac{2}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + \frac{2}{(m+1)^2(m+2)}$$

which gives

$$\mathbb{E}[D^2 f_X(Y)^2] = \frac{2}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{1}{m^3}\right)$$

so using (12), we obtain

$$\begin{aligned}
\sum_{k=0}^{m}\mathbb{E}[B_k^2] &= n + (n^2-n)\left((m+1)\mathbb{E}[D^2 f_X(Y)^2] + O\left(\frac{1}{m^{2\beta}}\right)\right) \\
&= n + (n^2-n)\left((m+1)\left(\frac{2}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{1}{m^3}\right)\right) + O\left(\frac{1}{m^{2\beta}}\right)\right) \\
&= n + (n^2-n)\left(\frac{2}{m+1}\mathbb{E}[f_X(Y)^2] + O\left(\frac{1}{(m+1)^2}\right) + O\left(\frac{1}{m^{2\beta}}\right)\right) \\
&= n + (n^2-n)\left(\frac{2}{m+1}\mathbb{E}[f_X(Y)^2] + O\left(\frac{1}{m^{2\beta}}\right)\right) \\
&= n + 2\frac{n^2-n}{m+1}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{m^{2\beta}}\right)
\end{aligned}$$

and finally,

$$\begin{aligned}
\mathbb{E}[B^2] &= \frac{1}{m+1}\sum_{k=0}^{m}\mathbb{E}[B_k^2] \\
&= \frac{n}{m+1} + 2\frac{n^2-n}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{m^{2\beta+1}}\right)
\end{aligned}$$

■

Corollary 15
Let $m = \Omega(n)$, then

$$\mathbb{E}[B^2] = \frac{n}{m} + 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + o\left(\frac{n^2}{m^2}\right)$$

Proof From the previous lemma:

$$\begin{aligned}\mathbb{E}[B^2] &= \frac{n}{m+1} + 2\frac{n^2-n}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{m^{2\beta+1}}\right)\\ &= \frac{n}{m} + 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + o\left(\frac{n^2}{m^2}\right)\end{aligned}$$

■

D.1.3. COMPUTING $\mathbb{P}(B$ IS ODD$)$

In this section, we prove Lemma 11. For this, we need Lemma 16, which itself uses the result of Lemma 17. Both are proven below in this section.

Lemma 11 *Let $\beta \in (0.5, 1)$, $m = \omega(n)$. Then, for n going to infinity,*

$$\mathbb{P}(B \textit{ is odd}) = \frac{n}{m} - 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)$$

Proof

B conditional on both Y and D follows a Binomial distribution of parameters $(n, P \triangleq p(Y, D))$, where $p(Y, D)$ is the probability of an item being in bin K, as defined in Definition 13.

The probability that a binomial random variable of parameters (n, p) is odd is $\frac{1}{2}(1-(1-2p)^n)$.

So

$$\begin{aligned}\mathbb{P}(B \text{ is odd}|Y, D) &= \frac{1}{2}(1-(1-2p(Y,D))^n)\\ &= \frac{1}{2}(1-(1-2P)^n)\end{aligned}$$

By linearity of the expectation

$$\mathbb{P}(B \text{ is odd}) = \mathbb{E}[\mathbb{P}(B \text{ is odd}|Y, D)] = \frac{1}{2}(1-\mathbb{E}[(1-2P)^n])$$

So, using Lemma 16 (below),

$$\begin{aligned}\mathbb{P}(B \text{ is odd}) &= \frac{1}{2}(1-(\mathbb{E}[(1-2P)^n]))\\ &= \frac{1}{2}\left(1-\left(1-2\frac{n}{m}+4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)\right)\right)\\ &= \frac{n}{m} - 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)\end{aligned}$$

■

We need the following result to prove Lemma 11:

Lemma 16 *Let $\beta \in (0.5, 1)$. Let $m = \omega(n)$ as n goes to infinity. Then we have*

$$\mathbb{E}[(1-2P)^n] = 1 - 2\frac{n}{m} + 4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)$$

Proof

Let $\beta \in (0.5, 1)$. Let $c_X \triangleq \max f_X$ and, for all $m \in \mathbb{N}$, let $c_m(\beta) \triangleq \pm\frac{2c}{m^{2\beta}}$, where c is the Lipschitz constant of f_X. To avoid clutter, we use the notation $c_m \triangleq c_m(\beta)$. We recall the definition of event $\mathcal{E}$:

$$\forall \beta \in (0,1), m \in \mathbb{N}, \quad \mathcal{E}(\beta, m) \triangleq \left(\bigcap_{k=0}^{m}\left(D_k \leq \frac{1}{m^\beta}\right)\right)$$

We will use the following structure to estimate $\mathbb{E}[(1-2P)^n]$:

$$\begin{aligned}
\mathbb{E}[(1-2P)^n] &\simeq \mathbb{E}[(1-2P)^n|\mathcal{E}] \\
&\simeq \mathbb{E}[(1-2\min(Df_X(Y) + \frac{c_m}{2}, 1))^n|\mathcal{E}] \\
&\simeq \mathbb{E}[(1-2\min(Df_X(Y) + \frac{c_m}{2}, 1))^n] \\
&\simeq \mathbb{E}[(1-2Df_X(Y) + c_m)^n] \\
&\simeq 1 - 2\frac{n}{m} + 4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2]
\end{aligned}$$

More precisely, we will prove the corresponding approximations:

1. $|\mathbb{E}[(1-2P)^n] - \mathbb{E}[(1-2P)^n|\mathcal{E}(\beta,m)]| \leq 2(m+1)e^{-m^{1-\beta}}$
2. $\mathbb{E}[(1-2P)^n|\mathcal{E}(\beta,m)]|$ is bounded by $\mathbb{E}[(1-2\min(Df_X(Y) \pm \frac{c_m}{2}, 1))^n|\mathcal{E}(\beta,m)]$
3. $|\mathbb{E}[(1-2\min(Df_X(Y)+\frac{c_m}{2},1))^n|\mathcal{E}(\beta,m)] - \mathbb{E}[(1-2\min(Df_X(Y)+\frac{c_m}{2},1))^n]| \leq 2(m+1)e^{-m^{1-\beta}}$
4. $|\mathbb{E}[(1-2\min(Df_X(Y)+\frac{c_m}{2},1))^n] - \mathbb{E}[(1-2Df_X(Y)+c_m)^n]| \leq e^{-m/c_X+o(m)}$
5. $\mathbb{E}[(1-2Df_X(Y)+c_m)^n] = 1 - 2\frac{n}{m} + 4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)$

Combining these 5 points directly yields the final result:

$$\mathbb{E}[(1-2P)^n] = 1 - 2\frac{n}{m} + 4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)$$

Proof of 5. 5. is immediate using Lemma 17 (below), by setting $v = 2c$, where c is the Lipschitz constant of f_X.

Proof of 1. and 3. We use Lemma 34 for both of them.

Let us set $W = (1-2P)^n$ and $Z = (1 - 2\min(Df_X(Y) + c_m, 1))^n$.

We have $|W| \leq 1$ and $|Z| \leq 1$. So, using Lemma 34, $|\mathbb{E}[W] - \mathbb{E}[W|\mathcal{E}]| \leq 2(1 - \mathbb{P}(\mathcal{E})) = 2(m+1)e^{-m^{\beta-1}}$ and $|\mathbb{E}[Z] - \mathbb{E}[Z|\mathcal{E}]| \leq 2(1 - \mathbb{P}(\mathcal{E})) = 2(m+1)e^{-m^{\beta-1}}$.

Proof of 4. For all $m, n \in \mathbb{N}$,

$$
\begin{aligned}
&\left|\mathbb{E}[(1 - 2\min(Df_X(Y) + \frac{c_m}{2}, 1))^n] - \mathbb{E}[(1 - 2(Df_X(Y) + \frac{c_m}{2}))^n]\right| \\
&= \left|\mathbb{E}\left[\mathbb{1}(Df_X(Y) + \frac{c_m}{2} \geq 1)\left(\mathbb{E}[(1 - 2\min(Df_X(Y) + \frac{c_m}{2}, 1))^n] - \mathbb{E}[(1 - 2Df_X(Y) + c_m)^n]\right)\right]\right| \\
&= \mathbb{P}(Df_X(Y) + \frac{c_m}{2} \geq 1)\left|\mathbb{E}[(1 - 2\min(Df_X(Y) + \frac{c_m}{2}, 1))^n] - (1 - 2Df_X(Y) + c_m)^n)|Df_X(Y) + \frac{c_m}{2} \geq 1]\right| \\
&= \mathbb{P}(Df_X(Y) + \frac{c_m}{2} \geq 1)\left|\mathbb{E}[(1-2)^n] - (1 - 2Df_X(Y) + c_m)^n)|Df_X(Y) + \frac{c_m}{2} \geq 1]\right| \\
&\leq \mathbb{P}\left(D \geq \frac{1 - \frac{c_m}{2}}{f_X(Y)}\right)\mathbb{E}[(2 + 2c_X + c_m)^n|Df_X(Y) + \frac{c_m}{2} \geq 1] \\
&\leq \mathbb{P}(D \geq \frac{1 - \frac{c_m}{2}}{c_X})(2 + 2c_X + c_m)^n \\
&\leq e^{-\frac{1-\frac{c_m}{2}}{c_X}m} e^{O(n)} \\
&\leq e^{-m/c_X + o(m)}
\end{aligned}
$$

where we used Lemma 23 with $\beta = 0$ and $a = \frac{1-\frac{c_m}{2}}{c_X}$ (D is a $Beta(1, m)$ because it is a mixture of $m+1$ different $Beta(1, m)$).

Proof of 2. f is upper bounded by c_X and $\mathcal{E}$ implies $D \leq \frac{c}{m^\beta}$ (Definition 14).
So, conditional on $\mathcal{E}$, for m large enough, we have $|2Df_X(Y)| + |\frac{2c}{m^{2\beta}}| \leq 1$, which implies $(1 - 2Df_X(Y) \pm \frac{2c}{m^{2\beta}}) \geq 0$.

We also have $0 \leq P \leq 1$, so $\mathcal{E}$ implies $|P - \min(Df_X(Y), 1)| \leq |P - Df_X(Y)| \leq \frac{c}{m^{2\beta}}$ (Lemma 25).

So, for m large enough, for all $n \in \mathbb{N}$, conditioned on $\mathcal{E}(\beta, m)$,

$$
\begin{aligned}
0 \leq Df_X(Y) - \frac{c}{m^{2\beta}} \quad &\leq P \quad \leq Df_X(Y) + \frac{c}{m^{2\beta}} \\
0 \leq \min(Df_X(Y) - \frac{c}{m^{2\beta}}, 1) \leq P &\leq \min(Df_X(Y) + \frac{c}{m^{2\beta}}, 1)
\end{aligned}
$$

So, for m large enough, for all $n \in \mathbb{N}$,

$$
\mathbb{E}\left[(1 - 2\min(Df_X(Y) - \frac{c}{m^{2\beta}}, 1))^n|\mathcal{E}\right] \geq \mathbb{E}\left[(1-2P)^n|\mathcal{E}\right] \geq \mathbb{E}\left[(1 - 2\min(Df_X(Y) + \frac{c}{m^{2\beta}}, 1))^n|\mathcal{E}\right]
$$

We conclude the proof of **2.** by recalling that we defined $c_m \triangleq \pm\frac{2c}{m^{2\beta}}$. ■

Lemma 17 *Let $c_m \triangleq \frac{v}{m^{2\beta}}$, where $\beta > 0.5$ and $v > 0$ are constants. Assume that $m = \omega(n)$. Then we have*

$$\mathbb{E}[(1 - 2Df_X(Y) + c_m)^n] = 1 - 2\frac{n}{m} + 4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)$$

Proof

We start by proving the result for $c_m = 0$, and we extend the analysis to $c_m \triangleq \frac{v}{m^{2\beta}}$ in a second part. We have

$$\mathbb{E}[(1 - 2Df_X(Y))^n] = 1 - 2n\mathbb{E}[f_X(Y)D] + 2n(n-1)\mathbb{E}[(f_X(Y)D)^2] + H \tag{18}$$

where $H \triangleq \sum_{k=3}^{n} \binom{n}{k}(-2)^k\mathbb{E}[(f_X(Y)D)^k]$ is the error term.
We will compute the two expectations, and upper bound $|H|$.
Lemmas 20 and 21 respectively give us the first and second moment of D conditioned on Y:

$$\mathbb{E}[D|Y = y > 0] = \frac{1}{m} - \frac{y^m}{m}, \quad \mathbb{E}[D|Y = 0] = \frac{1}{m+1}$$

$$\mathbb{E}[D^2|Y = y > 0] = \frac{2}{m(m+1)}(1-y^{m+1}-(m+1)(1-y)y^m), \quad \mathbb{E}[D^2|Y = 0] = \frac{2}{(m+1)(m+2)}$$

So, by the law of iterated expectations, we have the first expectation of (18):

$$\begin{aligned}
\mathbb{E}[f_X(Y)D] &= \mathbb{P}(Y > 0)\mathbb{E}[f_X(Y)D|Y > 0] + \mathbb{P}(Y = 0)\mathbb{E}[f_X(Y)D|Y = 0] \\
&= \frac{m}{m+1}\mathbb{E}\left[f_X(Y)\left(\frac{1}{m} - \frac{Y^m}{m}\right)\right] + \frac{1}{m+1}O\left(\frac{1}{m+1}\right) \\
&= \frac{m}{m+1}\left(\frac{1}{m}\mathbb{E}[f_X(Y)] - \frac{1}{m}\mathbb{E}[f_X(Y)Y^m]\right) + O\left(\frac{1}{m^2}\right) \\
&= \frac{1}{m} + O\left(\frac{1}{m^2}\right)
\end{aligned}$$

and the second expectation of (18):

$$\begin{aligned}
\mathbb{E}[(f_X(Y)D)^2] &= \mathbb{P}(Y > 0)\mathbb{E}[(f_X(Y)D)^2|Y > 0] + \mathbb{P}(Y = 0)\mathbb{E}[(f_X(Y)D)^2|Y = 0] \\
&= \frac{m}{m+1}\mathbb{E}\left[f_X(Y)^2\frac{2}{m(m+1)}(1 - Y^{m+1} - (m+1)(1-Y)Y^m)\right] + \frac{1}{m+1}O\left(\frac{2}{(m+1)(m+2)}\right) \\
&= \frac{2}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + \frac{2}{(m+1)^2}\mathbb{E}\left[f_X(Y)^2(Y^{m+1} - (m+1)(1-Y)Y^m))\right] + O\left(\frac{1}{m^3}\right) \\
&= \frac{2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{1}{m^3}\right)
\end{aligned}$$

Using these two results, we develop (18) to obtain

$$\mathbb{E}[(1-2Df(Y))^n] = 1 - 2n\left(\frac{1}{m} + O\left(\frac{1}{m^2}\right)\right) + 2n(n-1)\left(\frac{2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{1}{m^3}\right)\right) + H \tag{19}$$

$$= 1 - 2\frac{n}{m} + 4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + o\left(\frac{n^2}{m^2}\right) + H \tag{20}$$

Regarding the error term H, we have

$$|H| \triangleq \left|\sum_{k=3}^{n}\binom{n}{k}(-2)^k\mathbb{E}[(f_X(Y)D)^k]\right| \leq \sum_{k=3}^{n}\binom{n}{k}(2c_X)^k\mathbb{E}[D^k]$$

where c_X is the upper bound of f_X.
Because D is a $Beta(1,m)$, we have $\mathbb{E}[D^k] = \frac{k!m!}{(k+m)!} = \frac{1}{\binom{k+m}{m}}$
So

$$|H| \leq \sum_{k=3}^{n}\frac{\binom{n}{k}}{\binom{k+m}{k}}(2c_X)^k \tag{21}$$

Lemma 33 gives us $\frac{\binom{n}{k}}{\binom{k+m}{k}} \leq \frac{e^2}{2\pi}\left(\frac{n}{m}\right)^k$, so we have

$$\begin{aligned}
|H| &\leq \frac{e^2}{2\pi}\sum_{k=3}^{n}\left(\frac{n}{m}\right)^k(2c_X)^k \\
&\leq \frac{e^2}{2\pi}\sum_{k=3}^{\infty}\left(\frac{n}{m}2c_X\right)^k \\
&= \frac{e^2}{2\pi}\frac{\left(\frac{n}{m}2c_X\right)^3}{1-\frac{n}{m}2c_X} \\
&= O((n/m)^3) \\
&= o((n/m)^2)
\end{aligned}$$

where the last two equalities come from our assumption $m = \omega(n)$.

We replace H in (20) to obtain

$$\mathbb{E}[(1-2Df(Y))^n] = 1 - 2\frac{n}{m} + 4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + o\left(\frac{n^2}{m^2}\right) \tag{22}$$

We now compute the error created when putting the margin term c_m.
We will show that $\mathbb{E}[(1-2Df(Y)+c_m)^n] - \mathbb{E}[(1-2Df(Y))^n] = O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)$.

We have

$$
\begin{aligned}
&\mathbb{E}[(1-2Df(Y)+c_m)^n] - \mathbb{E}[(1-2Df(Y))^n] \\
&= nc_m + \frac{n(n-1)}{2}(c_m\mathbb{E}[2Df(Y)] + c_m^2) + H' - H \\
&= nc_m + \frac{n(n-1)}{2}\left(2c_m\left(\frac{1}{m} + O\left(\frac{1}{m^2}\right)\right) + c_m^2\right) + H' - H \\
&= O\left(\frac{n}{m^{2\beta}}\right) + \frac{n(n-1)}{2}\left(O\left(\frac{1}{m^{2\beta+1}}\right) + O\left(\frac{1}{m^{2(2\beta)}}\right)\right) + H' - H \\
&= O\left(\frac{n}{m^{2\beta}}\right) + H' - H
\end{aligned}
$$

where $H' = \sum_{k=3}^{n}\binom{n}{k}\mathbb{E}[(-2f_X(y)D + c_m)^k]$.

We know that $H = o\left(\frac{n^2}{m^2}\right)$, so we have

$$
\mathbb{E}[(1-2Df(Y)+c_m)^n] - \mathbb{E}[(1-2Df(Y))^n] = H' + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right) \tag{23}
$$

In addition, we have

$$
\begin{aligned}
|H'| &\leq \sum_{k=3}^{n}\binom{n}{k}\mathbb{E}[(2f_X(Y)D + c_m)^k] \\
&\leq \sum_{k=3}^{n}\binom{n}{k}\sum_{j=0}^{k}\binom{k}{j}(2c_X)^j\mathbb{E}[D^j]c_m^{k-j} \\
&= \sum_{k=3}^{n}\binom{n}{k}\left((2c_X)^k\mathbb{E}[D^k] + \sum_{j=0}^{k-1}\binom{k}{j}(2c_X)^j\mathbb{E}[D^j]c_m^{k-j}\right) \\
&= o\left(\frac{n^2}{m^2}\right) + \sum_{k=3}^{n}\binom{n}{k}\sum_{j=0}^{k-1}\binom{k}{j}(2c_X)^j\mathbb{E}[D^j]c_m^{k-j} \\
&= o\left(\frac{n^2}{m^2}\right) + \sum_{k=3}^{n}\binom{n}{k}\sum_{j=0}^{k-1}\binom{k}{j}(2c_X)^j\frac{1}{\binom{j+m}{m}}\left(\frac{2c}{m^{2\beta}}\right)^{k-j} \\
&\leq o\left(\frac{n^2}{m^2}\right) + \sum_{k=3}^{n}\binom{n}{k}\sum_{j=0}^{k-1}\frac{\binom{k}{j}}{\binom{j+m}{m}}\left(\frac{2c}{m^{2\beta-1}}\right)^{k-j}\frac{1}{m^{k-j}}(2c_X)^j
\end{aligned}
$$

Using $\frac{\binom{k}{j}}{\binom{j+m}{m}} \leq \frac{e^2}{\pi}\left(\frac{k}{m}\right)^j$ from Lemma 33 and for m large enough (such that $m^{2\beta-1} \geq 2c$), we have

$$
\begin{aligned}
|H'| &\leq o\left(\frac{n^2}{m^2}\right) + \sum_{k=3}^{n}\binom{n}{k}\sum_{j=0}^{k-1}\left(\frac{k}{m}\right)^j \frac{1}{m^{k-j}}(2c_X)^j \\
&\leq o\left(\frac{n^2}{m^2}\right) + \sum_{k=3}^{n}\binom{n}{k}\frac{1}{m^k}\sum_{j=0}^{k-1}k^j(2c_X)^j \\
&\leq o\left(\frac{n^2}{m^2}\right) + \sum_{k=3}^{n}\left(\frac{ne}{k}\right)^k\frac{1}{m^k}\sum_{j=0}^{k-1}k^j(2c_X)^j \\
&\leq o\left(\frac{n^2}{m^2}\right) + \sum_{k=3}^{n}\left(\frac{ne}{m}\right)^k\frac{1}{k^k}\frac{k^k(2c_X)^k-1}{k(2c_X)-1} \\
&\leq o\left(\frac{n^2}{m^2}\right) + \sum_{k=3}^{n}\left(\frac{ne(2c_X)}{m}\right)^k \\
&= o\left(\frac{n^2}{m^2}\right)
\end{aligned}
$$

where we used the general inequality $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$.

Using this in (23), we get that $\mathbb{E}[(1-2Df(Y)+c_m)^n] - \mathbb{E}[(1-2Df(Y))^n] = O(\frac{n}{m^{2\beta}}) + o(\frac{n^2}{m^2})$, so from (22), we obtain our final result:

$$
\mathbb{E}[(1-2Df(Y)+c_m)^n] = 1 - 2\frac{n}{m} + 4\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)
$$

■

D.2. Proof of Theorem 5

Lemma 7 gives us $\mathbb{E}[F] = \frac{1}{2}(m+1)(\mathbb{E}[B^2] - \mathbb{P}(B \text{ is odd}))$. In this expression, the last term is at most one, so it can be ignored if $\mathbb{E}[B^2]$ goes to infinity. This is the case for Theorem 5. Theorem 6 treats the case of a constant MSF, for which we need this second term.

Theorem 5 *[For $f_Y = 1$] Assume that there exists $r \in \mathbb{R}^+$ s.t. $m \sim rn$ as n goes to infinity, then*

$$
\left|\mathbb{E}[F] - n\left(\frac{1}{2} + \frac{1}{r}\left[f_X(Y)^2\right]\right)\right| \leq \frac{r}{2}n + o(n)
$$

Proof

Lemma 7 gives $\mathbb{E}[F] = \frac{1}{2}(m+1)(\mathbb{E}[B^2] - \mathbb{P}(B \text{ is odd}))$ and we have $1 \geq \mathbb{P}(B \text{ is odd})) \geq 0$, so $\frac{m+1}{2}(\mathbb{E}[B^2]-1) \leq \mathbb{E}[F] \leq \frac{m+1}{2}\mathbb{E}[B^2]$ and

$$
-\frac{m+1}{2} \leq \mathbb{E}[F] - \frac{m+1}{2}\mathbb{E}[B^2] \leq 0
$$

From Lemma 10, we have

$$\forall \beta \in (0.5, 1), \quad \mathbb{E}[B^2] = \frac{n}{m+1} + 2\frac{n^2-n}{(m+1)^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{m^{2\beta+1}}\right)$$

So, put back in the previous equation,

$$-\frac{m+1}{2} \leq \mathbb{E}[F] - \left(\frac{1}{2}n + \frac{n^2-n}{m+1}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{m^{2\beta}}\right)\right) \qquad \leq 0 \qquad (24)$$

We have $\exists r \in \mathbb{R}^+$ *s.t.* $m \sim rn$, so $m = rn + o(n)$.
Focusing on the middle term of equation (24),

$$\begin{aligned}
&\frac{1}{2}n + \frac{n^2-n}{m}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{m^{2\beta}}\right) \\
&= \frac{1}{2}n + \frac{n^2-n}{(rn+o(n))}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n^2}{(rn+o(n))^{2\beta}}\right) \\
&= \frac{1}{2}n + \frac{n^2-n}{rn}(1+o(1))\mathbb{E}[f_X(Y)^2] + O(n^{2-2\beta}) \\
&= \frac{1}{2}n + \frac{n-1}{r}\mathbb{E}[f_X(Y)^2] + O(n^{2(1-\beta)}) \\
&= n\left(\frac{1}{2} + \frac{1}{r}\mathbb{E}[f_X(Y)^2]\right) + o(n)
\end{aligned}$$

Putting the result back in equation (24), we have
$-\frac{rn+o(n)}{2} \leq n\left(\frac{1}{2} + \frac{1}{r}\mathbb{E}[f_X(Y)^2]\right) + o(n) \leq 0$, which implies

$$\left|\mathbb{E}[F] - n\left(\frac{1}{2} + \frac{1}{r}\mathbb{E}[f_X(Y)^2]\right)\right| \leq \frac{r}{2}n + o(n)$$

■

This version of Theorem 5 depends on the *rescaled* density f_X, *i.e.* under the assumption that $f_Y = 1$. In order to generalize to a general f_Y, we use Lemma 26 (Appendix G). It states that if (f_X, f_Y) are the true densities of the scores and thresholds, and $(f_{X'}, 1)$ are their rescaled counterparts, then $\mathbb{E}[f_{X'}(Y')^2] = \mathbb{E}\left[\left(\frac{f_X(Y)}{f_Y(Y)}\right)^2\right]$. This gives us the final version of the theorem:

Theorem 5 *[For any f_Y] Assume that there exists $r \in \mathbb{R}^+$* s.t. *$m \sim rn$ as n goes to infinity, then*

$$\left|\mathbb{E}[F] - n\left(\frac{1}{2} + \frac{1}{r}\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right]\right)\right| \leq \frac{r}{2}n + o(n)$$

D.3. Proof of Theorem 6

We provide here the proof of Theorem 6, which uses Lemmas 7, 10 (Corollary 15) and 11.

Theorem 6 *[For $f_Y = 1$] Assume that there exists $r \in \mathbb{R}^+, \gamma > 1$* s.t. *$m \sim rn^\gamma$, as n goes to infinity, then*

$$\mathbb{E}[F] \sim \frac{2}{r}n^{2-\gamma}\mathbb{E}[f_X(Y)^2]$$

Proof

Let us set $m = rn^\gamma + o(n^\gamma)$, $\gamma > 1$

Using Corollary 15 and Lemma 11, for all $\beta \in (0.5, 1)$, we have
Corollary 15:

$$\mathbb{E}[B^2] = \frac{n}{m} + 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + o\left(\frac{n^2}{m^2}\right)$$

Then Lemma 11:

$$\mathbb{P}(B \text{ is odd}) = \frac{n}{m} - 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)$$

So, using Lemma 7, we have

$$\mathbb{E}[F] = \frac{(m+1)}{2}(\mathbb{E}[B^2] - \mathbb{P}(B \text{ is odd})) \tag{25}$$

$$= \frac{(m+1)}{2}\left(\frac{n}{m} + 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] - \frac{n}{m} + 2\frac{n^2}{m^2}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta}}\right) + o\left(\frac{n^2}{m^2}\right)\right) \tag{26}$$

$$= 2\frac{n^2}{m}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^{2\beta-1}}\right) + o\left(\frac{n^2}{m}\right) \tag{27}$$

$$\tag{28}$$

By reparametrizing, for any $\alpha < 1$, we have

$$\mathbb{E}[F] = 2\frac{n^2}{m}\mathbb{E}[f_X(Y)^2] + O\left(\frac{n}{m^\alpha}\right) + o\left(\frac{n^2}{m}\right)$$

If it were true for $\alpha = 1$, we would have $\mathbb{E}[F] = 2\frac{n^2}{m}\mathbb{E}[f_X(Y)^2] + o\left(\frac{n^2}{m}\right)$ as long $m = \omega(n)$.
However, here we have to pick the value of α close enough to 1, depending on the value of γ.
$O\left(\frac{n}{m^\alpha}\right) = O\left(\frac{n}{(rn^\gamma + o(n^\gamma))^\alpha}\right) = O\left(n^{1-\alpha\gamma}\right)$ and $o\left(\frac{n^2}{m}\right) = o\left(n^{2-\gamma}\right)$
So by setting for instance $\alpha = 1 - \frac{1}{2\gamma}$, we have $O\left(\frac{n}{m^\alpha}\right) = o\left(\frac{n^2}{m}\right)$
and

$$\mathbb{E}[F] = 2\frac{n^2}{m}\mathbb{E}[f_X(Y)^2] + o\left(\frac{n^2}{m}\right)$$

i.e.

$$\mathbb{E}[F] \sim 2\frac{n^2}{m}\mathbb{E}[f_X(Y)^2]$$

and

$$\mathbb{E}[F] \sim \frac{2}{r}n^{2-\gamma}\mathbb{E}[f_X(Y)^2] \tag{29}$$

■

As we did for Theorem 5, we generalize to a general f_Y using Lemma 26, thus obtaining the final form of Theorem 6.

Theorem 6 *[For general f_Y] Assume that there exists $r \in \mathbb{R}^+, \gamma > 1$ s.t. $m \sim rn^\gamma$, as n goes to infinity, then*

$$\mathbb{E}[F] \sim \frac{2}{r} n^{2-\gamma} \mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right]$$

Appendix E. Properties of Random Variables

In this section, we present results on the random variables which appear in our proofs of the previous section.

Lemma 18

$$\sum_{k=0}^{m} \mathbb{E}[B_k^2] = n + (n^2 - n) \sum_{k=0}^{m} \mathbb{E}[P_k^2]$$

Proof For all $k \in [m]$, we have that B_k conditioned on (D_k, Y_k) follows a binomial distribution of parameters (P_k, n), with $P_k \triangleq \int_{Y_k}^{Y_k + D_k} f_X(x)dx$. So, by the tower rule,

$$\begin{aligned}
\sum_{k=0}^{m} \mathbb{E}[B_k^2] &= \sum_{k=0}^{m} \mathbb{E}[\mathbb{E}[B_k^2 | D_k, Y_k]] \\
&= \sum_{k=0}^{m} \mathbb{E}[nP_k + (n^2 - n)P_k^2] \\
&= \sum_{k=0}^{m} n\mathbb{E}[P_k] + (n^2 - n) \sum_{k=0}^{m} \mathbb{E}[P_k^2] \\
&= n\mathbb{E}[\sum_{k=0}^{m} P_k] + (n^2 - n) \sum_{k=0}^{m} \mathbb{E}[P_k^2] \\
&= n\mathbb{E}[1] + (n^2 - n) \sum_{k=0}^{m} \mathbb{E}[P_k^2] \\
&= n + (n^2 - n) \sum_{k=0}^{m} \mathbb{E}[P_k^2]
\end{aligned}$$

■

Lemma 12 $\forall \beta \in (0.5, 1)$, *when m goes to infinity,*

$$\mathbb{E}\left[\sum_{k=0}^{m} P_k^2\right] = \sum_{k=0}^{m} \mathbb{E}[(D_k f_X(Y_k))^2] + O\left(\frac{1}{m^{2\beta}}\right)$$

Proof

Lemma 25 gives us

$$\mathcal{E} \Rightarrow \forall k \in [m], |P_k - D_k f_X(Y_k)| \leq \frac{c}{m^{2\beta}} \triangleq g_m \tag{30}$$

We have assumed $\beta > 0.5$, which implies $g_m = o\left(\frac{1}{m}\right)$.
Then, for all m, conditioned on $\mathcal{E}$, we have

$$\begin{aligned}
D_k f_X(Y_k) - g_m \leq P_k &\leq D_k f_X(Y_k) + g_m \\
P_k D_k f_X(Y_k) - P_k g_m \leq P_k^2 &\leq P_k D_k f_X(Y_k) + P_k g_m \\
\sum_{k=0}^{m}(D_k f_X(Y_k) - g_m) D_k f_X(Y_k) - g_m \leq \sum_{k=0}^{m} P_k^2 &\leq \sum_{k=0}^{m}(D_k f_X(Y_k) + g_m) D_k f_X(Y_k) + g_m \\
\sum_{k=0}^{m}(D_k f_X(Y_k))^2 - g_m \sum_{k=0}^{m} D_k f_X(Y_k) - g_m \leq \sum_{k=0}^{m} P_k^2 &\leq \sum_{k=0}^{m}(D_k f_X(Y_k))^2 + g_m \sum_{k=0}^{m} D_k f_X(Y_k) + g_m \\
\sum_{k=0}^{m}(D_k f_X(Y_k))^2 - g_m \sum_{k=0}^{m}(P_k + g_m) - g_m \leq \sum_{k=0}^{m} P_k^2 &\leq \sum_{k=0}^{m}(D_k f_X(Y_k))^2 + g_m \sum_{k=0}^{m}(P_k + g_m) + g_m \\
\sum_{k=0}^{m}(D_k f_X(Y_k))^2 - g_m(1 + m g_m) - g_m \leq \sum_{k=0}^{m} P_k^2 &\leq \sum_{k=0}^{m}(D_k f_X(Y_k))^2 + g_m(1 + m g_m) + g_m
\end{aligned}$$

where we used equation (30) several times and the property $\sum_{k=0}^{m} P_k = 1$.

So, for all m, conditional on $\mathcal{E}$, we have

$$-g_m(2 + m g_m) \leq \sum_{k=0}^{m} P_k^2 - \sum_{k=0}^{m}(D_k f_X(Y_k))^2 \leq g_m(2 + m g_m)$$

which implies

$$-g_m(2 + m g_m) \leq \mathbb{E}[\sum_{k=0}^{m} P_k^2 | \mathcal{E}] - \mathbb{E}[\sum_{k=0}^{m}(D_k f_X(Y_k))^2 | \mathcal{E}] \leq g_m(2 + m g_m)$$

So, using Lemma 34 to remove the conditioning, and letting $c_X = \max(f_X)$

$$\begin{aligned}
-g_m(2 + m g_m) - (1 - \mathbb{P}(\mathcal{E})) - (m+1)(1 - \mathbb{P}(\mathcal{E}))c_X^2 &\leq \mathbb{E}[\sum_{k=0}^{m} P_k^2] - \mathbb{E}[\sum_{k=0}^{m}(D_k f_X(Y_k))^2] \\
\mathbb{E}[\sum_{k=0}^{m} P_k^2] - \mathbb{E}[\sum_{k=0}^{m}(D_k f_X(Y_k))^2] &\leq g_m(2 + m g_m) + (1 - \mathbb{P}(\mathcal{E})) + (m+1)(1 - \mathbb{P}(\mathcal{E}))c_X^2
\end{aligned}$$

This gives us

$$|\mathbb{E}[\sum_{k=0}^{m} P_k^2] - \mathbb{E}[\sum_{k=0}^{m}(D_k f_X(Y_k))^2]| \leq g_m(2+mg_m) + (1-\mathbb{P}(\mathcal{E})) + (m+1)(1-\mathbb{P}(\mathcal{E}))c_X^2$$
$$= O(g_m)$$
$$= O\left(\frac{1}{m^{2\beta}}\right)$$

Because, from Lemma 24, we have $1-\mathbb{P}(\mathcal{E}) \leq (m+1)e^{-m^{1-\beta}}$

■

Lemma 12 makes the expression $\mathbb{E}[(D_k f_X(Y_k))^2]$ appear. For all k, the left extremity (Y_k) of the bin and the length (D_k) of the bin are not independent random variables. For this reason, we will need information on their joint distribution. For this, we derive the conditional distribution of D given Y and its first two moments in Lemmas 19, 20, 21.

Lemma 19 *For all $y \in (0,1]$, the conditional density of D given $Y = y$ is*

$$f_{D|Y=y>0}(x) = (m-1)(1-x)^{m-2}\mathbb{1}(x < 1-y) + y^{m-1}\delta_{1-y}(x)$$

For the case $y = 0$, we have

$$f_{D|Y=0}(x) = m(1-x)^{m-1}$$

Proof Let us start by the case $y = 0$. Conditional on $Y_K = 0$, we know that $K = 0$, so $D \triangleq D_K = D_0$. D_0 is a $Beta(1,m)$, and is independent from K. So $D|Y = 0$ is a $Beta(1,m)$. *i.e.* $f_{D|Y=0}(x) = m(1-x)^{m-1}$.

We now consider $y \in (0,1]$. Let $d \in (0,1]$.
Then the probability $\mathbb{P}(D > d|Y = y)$ is zero if $d \geq 1-y$ (the bin cannot be longer than the space between its left extremity and 1).
If $d < 1-y$, the event $(D > d)$ is equivalent to having no threshold in the interval $[y, y+d]$ So $\mathbb{P}(D > d|Y = y)$ is the probability of having no threshold in $[y, y+d]$ given that there is a threshold at y. The sampling of the thresholds is *iid* uniform in $[0,1]$. Therefore, the probability for one threshold to not be in the interval is $(1-d)$, and the probability for no threshold to be there is $(1-d)^{m-1}$ (one threshold is fixed at position y by the conditioning). So, we have:

For $y > 0$, $\mathbb{P}(D > d|Y = y) = (1-d)^{m-1}\mathbb{1}(d < 1-y)$

By differentiating the ccdf, we obtain the result:

$$f_{D|Y=y}(x) = (m-1)(1-x)^{m-2}\mathbb{1}(x < 1-y) + y^{m-1}\delta_{1-y}(x)$$

■

Lemma 20

$$\forall m \in \mathbb{N}, \quad \mathbb{E}[D|Y=0] = \frac{1}{m+1}$$
$$\forall m \geq 1, \quad \mathbb{E}[D|Y=y>0] = \frac{1}{m} - \frac{y^m}{m}$$

Proof

Recall that by definition, $D = D_K$, where K is uniform on $[m] \cup \{0\}$ and that $Y_0 = 0$ by convention. This means that $Y_K = 0$ implies $K = 0$, ie $D_K = D_0$, which follows a $Beta(1, m)$ (Lemma 22).

So

$$\mathbb{E}[D|Y = 0] = \frac{1}{m+1}$$

In addition, using Lemma 19,

$$\begin{aligned}
\mathbb{E}[D^k|Y = y > 0] &= \int_0^1 x^k (m-1)(1-x)^{m-2}(\mathbb{1}(x < 1-y) + y^{m-1}\delta_{1-y}(x))dx \\
&= (m-1)\int_0^{1-y} x^k(1-x)^{m-2}dx + y^{m-1}\int_0^1 x^k\delta_{1-y}(x))dx \\
&= (m-1)\int_0^{1-y} x^k(1-x)^{m-2}dx + y^{m-1}(1-y)^k
\end{aligned}$$

In particular, for $k = 1$, using integration by parts,

$$\begin{aligned}
\mathbb{E}[D|Y = y > 0] &= (m-1)\int_0^{1-y} x(1-x)^{m-2}dx + y^{m-1}(1-y) \\
&= (m-1)\left[x\frac{-(1-x)^{m-1}}{m-1}\right]_0^{1-y} - (m-1)\int_0^{1-y}\frac{-(1-x)^{m-1}}{m-1}dx + y^{m-1}(1-y) \\
&= -(1-y)y^{m-1} - \frac{y^m}{m} + \frac{1}{m} + y^{m-1}(1-y) \\
&= \frac{1}{m} - \frac{y^m}{m}
\end{aligned}$$

■

Lemma 21

$$\forall m \in \mathbb{N}, \quad \mathbb{E}[D^2|Y = 0] = \frac{2}{(m+1)(m+2)}$$

$$\forall m \geq 1, \quad \mathbb{E}[D^2|Y = y > 0] = \frac{2}{m(m+1)}\left(1 - y^{(m+1)} - (m+1)y^m(1-y)\right)$$

Proof

Recall that by definition, $D = D_K$, where K is uniform on $[m] \cup \{0\}$ and that $Y_0 = 0$ by convention. This means that $Y_K = 0$ implies $K = 0$, ie $D_K = D_0$, which follows a $Beta(1, m)$ (Lemma 22). So

$$\mathbb{E}[D_K^2|Y_K = 0] = \mathbb{E}[D_0^2] = \frac{2}{(m+1)(m+2)}$$

We now treat the case $Y = y > 0$. We can prove it in two different ways.

Method 1 :

Lemma 19 gives us $f_{D|Y=y>0}(x) = (m-1)(1-x)^{m-2}\mathbb{1}(x < 1-y) + y^{m-1}\delta_{1-y}(x)$. So

$$\begin{aligned}
\mathbb{E}[D^2|Y=y] &= \int_0^1 x^2 f_{D|Y=y}(x)dx \\
&= \int_0^1 x^2((m-1)(1-x)^{m-2}\mathbb{1}(x<1-y) + y^{m-1}\delta_{1-y}(x))dx \\
&= (m-1)\int_0^{1-y} x^2(1-x)^{m-2}dx + \int_0^1 x^2 y^{m-1}\delta_{1-y}(x)dx \\
&= (m-1)\int_0^{1-y} x^2(1-x)^{m-2}dx + (1-y)^2 y^{m-1}
\end{aligned}$$

with

$$\begin{aligned}
\int_0^{1-y} x^2(1-x)^{m-2}dx &= \left[x^2\frac{-(1-x)^{m-1}}{m-1}\right]_0^{1-y} - \int_0^{1-y} 2x\frac{-(1-x)^{m-1}}{m-1}dx \\
&= \left[x^2\frac{-(1-x)^{m-1}}{m-1}\right]_0^{1-y} - 2\left[x\frac{(1-x)^m}{(m-1)m}\right]_0^{1-y} - 2\int_0^{1-y}\frac{-(1-x)^m}{(m-1)m}dx \\
&= -(1-y)^2\frac{y^{m-1}}{m-1} - 2(1-y)\frac{y^m}{(m-1)m} - 2\frac{y^{m+1}-1}{(m-1)m(m+1)}
\end{aligned}$$

so finally

$$\begin{aligned}
\mathbb{E}[D^2|Y=y] &= -2(1-y)\frac{y^m}{m} + 2\frac{1-y^{m+1}}{m(m+1)} \\
&= \frac{2}{m(m+1)}\left(-(m+1)(1-y)y^m + 1 - y^{m+1}\right) \\
&= \frac{2}{m(m+1)}\left(1 - y^{m+1} - (m+1)(1-y)y^m\right)
\end{aligned}$$

Method 2 :

The thresholds are sampled *iid* uniformly in $[0,1]$ (because $f_Y = 1$). In particular, each one of the (unordered) thresholds has a probability y of being smaller than y, independently of each other. We have $m-1$ such thresholds, because there are a total of m thresholds, including $Y_K = y$ itself. Therefore, the distribution of K conditional on $Y_K = y \neq 0$ is one plus a binomial distribution of parameters $(y, m-1)$.
ie $\forall 0 < y \leq 1, \quad K-1|Y_K = y \sim Bin(y, m-1)$.

In addition, conditioned on $K = k$, we know that we have exactly $m + 1 - k$ thresholds greater than Y_k, and that the distribution of these thresholds is uniform *iid* on $[Y_k, 1]$. This means that the distance $D_k = Y_{k+1} - Y_k$ follows a Beta distribution rescaled on the interval $[Y_k, 1]$.

So, formally, $\frac{D_k}{1-y}|Y_k = y \sim Beta(1, m-k)$ (with the convention $Beta(1,0)$ is a Dirac in 1).

In general, if $Z \sim Beta(1, a)$, then $\mathbb{E}[Z^2] = \frac{2}{(1+a)(2+a)}$. This gives us

$$\mathbb{E}[D_K^2|Y_K = y, K] = 2(1-y)^2\mathbb{E}\left[\frac{1}{(1+m-K)(2+m-K)}\middle|Y_K = y, K\right]$$

$$\begin{aligned}
\mathbb{E}[D_K^2|Y_K = y] &= \mathbb{E}[\mathbb{E}[D_K^2|Y_K = y, K]] \\
&= 2(1-y)^2\mathbb{E}\left[\frac{1}{(1+m-K)(2+m-K)}\middle|Y_K = y\right] \\
&= 2(1-y)^2\sum_{k=1}^{m}\frac{1}{(1+m-k)(2+m-k)}\mathbb{P}(K = k|Y_K = y) \\
&= 2(1-y)^2\sum_{k=1}^{m}\frac{1}{(1+m-k)(2+m-k)}\binom{m-1}{k-1}y^{k-1}(1-y)^{m-k} \\
&= 2\sum_{k=1}^{m}\frac{\frac{1}{(m+1)m}(m+1)!}{(k-1)!(m-k+2)!}y^{k-1}(1-y)^{m-k+2} \\
&= \frac{2}{(m+1)m}\sum_{k=1}^{m}\binom{m+1}{k-1}y^{k-1}(1-y)^{m-k+2} \\
&= \frac{2}{(m+1)m}\sum_{k=0}^{m-1}\binom{m+1}{k}y^{k}(1-y)^{m-k+1} \\
&= \frac{2}{(m+1)m}\left(\left(\sum_{k=0}^{m+1}\binom{m+1}{k}y^k(1-y)^{m+1-k}\right) - y^{m+1} - (m+1)y^m(1-y)\right) \\
&= \frac{2}{(m+1)m}\left(1 - y^{m+1} - (m+1)y^m(1-y)\right)
\end{aligned}$$

■

Because we made the assumption that $f_Y = 1$, the user thresholds are *iid* uniform random variables. The following lemma shows that for all k, the bin length D_k follows a $Beta(1, m)$.

Lemma 22 *Let $Z_{i\in[m]}$ be iid uniform random variables on $[0, 1]$. Let $Z_{(i),i\in[m]}$ be the order statistics of Z. Then*
$\forall i \in [m-1]$, $Z_{(i+1)} - Z_{(i)} \sim Beta(1, m)$, *i.e.* $f_{Z_{(i+1)}-Z_{(i)}}(x) = m(1-x)^{m-1}\mathbb{1}(x \in [0, 1])$

Proof In order to generate *iid* uniform random variables on $[0, 1]$, one can sample $m + 1$ uniform random variables on a unit circle and then pick one of them to be the start of the $[0, 1]$ interval (the 0 point). This way, it is clear that $Z_{(1)} - 0$ has the same distribution as $Z_{(i+1)} - Z_{(i)}$ for all $i \in [m-1]$. It is well-known that the minimum of m uniform random variables on $[0, 1]$ follows a $Beta(1, m)$,

so we also have $Z_{(i+1)} - Z_{(i)} \sim Beta(1, m)$.
(Proof adapted from the one given by Liran Katzir on Stack Exchange) ■

Lemma 23 *Let $m \in \mathbb{N}$, let $Z \sim Beta(1, m)$. For all $\beta \in [0, 1], a > 0$,*

$$\mathbb{P}\left(Z \leq \frac{a}{m^\beta}\right) \geq 1 - e^{-am^{1-\beta}}$$

Proof

$$\mathbb{P}(Z \geq \frac{a}{m^\beta}) = \int_{\frac{a}{m^\beta}}^{1} m(1-x)^m dx = (1 - \frac{a}{m^\beta})^m$$

Using the fact that $\forall x \in \mathbb{R}, \log(1-x) \leq -x$,

$$\log\left((1 - \frac{a}{m^\beta})^m\right) = m \log\left(1 - \frac{a}{m^\beta}\right) \leq -m\frac{a}{m^\beta} = -am^{1-\beta}$$

which implies $\mathbb{P}(Z \geq \frac{a}{m^\beta}) = (1 - \frac{a}{m^\beta})^m \leq e^{-am^{1-\beta}}$. ■

Appendix F. Uniform upper bound on the length of the bins

Because we assume f_X to be c-Lipschitz, we understand that, when m grows large, f_X will be almost constant over each bin interval $[Y_k, Y_{k+1}]$. In this section, we formalize this intuition. For this, we define a probabilistic event $\mathcal{E}$ under which the length of all bins is upper bounded. In Lemma 24, we show that this event happens with probability going exponentially to 1.

Definition 14 *For all $\beta \in (0, 1), m \in \mathbb{N}$, we define the event $\mathcal{E}(\beta, m)$ as follows:*

$$\mathcal{E}(\beta, m) \triangleq \left(\bigcap_{k=0}^{m} \left(D_k \leq \frac{1}{m^\beta}\right)\right)$$

Lemma 24

(i) $\forall \beta \in (0, 1), \forall m \in \mathbb{N}$

$$\mathbb{P}\left(\mathcal{E}(\beta, m)\right) \geq 1 - (m+1)e^{-m^{1-\beta}}$$

(ii) $\mathcal{E}(\beta, m)$ implies that the variation of f_X inside any bin is smaller than $\frac{c}{m^\beta}$, ie

$$\mathcal{E}(\beta, m) \Rightarrow \left(\bigcap_{k=0}^{m} \left(\forall x \in [Y_k, Y_{k+1}], |f_X(x) - f_X(Y_k)| \leq \frac{c}{m^\beta}\right)\right)$$
$$= \left(\sup_{k \in [m]} \sup_{x \in [Y_k, Y_{k+1}]} |f_X(x) - f_X(Y_k)| \leq \frac{c}{m^\beta}\right)$$

where c is the Lipschitz constant of f_X.

Proof (i) Let $m \in \mathbb{N}$. Because the thresholds are *iid* uniform on $[0,1]$, we have that $\forall k \in [m] \cup \{0\}, D_k \sim Beta(1,m)$ (this is a property of the order statistics of the uniform distribution, see Lemma 22 below in Appendix).

Lemma 23 gives us

$$\forall k \in [m], \mathbb{P}(D_k \leq \frac{1}{m^\beta}) \geq 1 - e^{-m^{1-\beta}}$$

So, by union bound on all k in $\mathbb{N}$,

$$\begin{aligned}\mathbb{P}(\forall k \in [m], D_k \leq \frac{1}{m^\beta}) &\geq 1 - \sum_{k=0}^{m} e^{-m^{1-\beta}} \\ &= 1 - (m+1)e^{-m^{1-\beta}}\end{aligned}$$

(ii) By definition, $\mathcal{E}(\beta,m) \Rightarrow \left(\forall k \in [m], D_k \leq \frac{1}{m^\beta}\right)$.

Coupled with the fact that f_X is c-Lipschitz, we obtain

$$\forall k \in [m], \forall x \in [Y_k, Y_{k+1}], |f_X(x) - f_X(Y_k)| \leq cD_k$$

So by combining the two, we have

$$\mathcal{E}(\beta,m) \Rightarrow \left(\forall k \in [m], \forall x \in [Y_k, Y_{k+1}], |f_X(x) - f_X(Y_k)| \leq \frac{c}{m^\beta}\right)$$

■

Because the density f_X of the items scores is not constant on $[0,1]$, the conditional probability of an item being in bin k conditioned the user thresholds does not depend only on the length D_k of the bin, but also on its position in the $[0,1]$ interval.
We first provide some notation for this conditional probability.

We now provide an approximation of P_k, by conditioning on event $\mathcal{E}$.

Lemma 25 *For all $i \in [n]$, the (conditional) probability of item i being in bin k can be approximated by $D_k f_X(Y_k)$.*

$\forall \in \mathbb{N}, \forall \beta \in (0,1)$,

$$\mathcal{E}(\beta,m) \Rightarrow \left(\forall k \in [m], \quad |P_k - D_k f_X(Y_k)| \leq \frac{c}{m^{2\beta}}\right)$$

where c is the Lipschitz constant of f_X.

Proof Using Lemma 24, conditional on the event $\mathcal{E}(\beta,m)$, we have, for all $k \in [m] \cup \{0\}$

$$\begin{aligned}\forall x \in [Y_k, Y_{k+1}], \quad f_X(Y_k) - \frac{c}{m^\beta} \leq f_X(x) &\leq f_X(Y_k) + \frac{c}{m^\beta} \\ D_k(f_X(Y_k) - \frac{c}{m^\beta}) \leq \int_{Y_k}^{Y_k+D_k} f_X(x)dx &\leq D_k(f_X(Y_k) + \frac{c}{m^\beta}) \\ D_k(f_X(Y_k) - \frac{c}{m^\beta}) \leq P_k &\leq D_k(f_X(Y_k) + \frac{c}{m^\beta}) \\ -D_k\frac{c}{m^\beta} \leq P_k - D_k f_X(Y_k) &\leq D_k\frac{c}{m^\beta}\end{aligned}$$

and $\mathcal{E}(\beta, m) \Rightarrow D_k \leq \frac{1}{m^\beta}$, so

$$\mathcal{E}(\beta, m) \Rightarrow |P_k - D_k f_X(Y_k)| \leq D_k \frac{c}{m^\beta} \leq \frac{c}{m^{2\beta}}$$

■

The probability of event $\mathcal{E}$ goes to 1 as m goes to infinity (Lemma 24). Therefore, it seems intuitive that an expectation conditional on this event should not be too different from the unconditional expectation. We now prove a lemma that formalizes this intuition. This result will be useful to approximate unconditional expectations by using properties valid under $\mathcal{E}$.

Appendix G. Influence of f_X and f_Y

G.1. Reducing the problem to the $f_Y = 1$ case

As explained in Section 3.1, we can assume without loss of generality that $f_Y = 1$. We formally prove it in this section. We start by recalling the notation defined in Appendix B, but without the assumption $f_Y = 1$. Then, we show that we can rescale everything in order to obtain an equivalent setting in which we have $f_Y = 1$.

- f_X: the density of the item scores on $[0, 1]$. Scores are *iid* on the interval.
- f_Y: the density of the user thresholds on $[0, 1]$. Thresholds are *iid* on the interval.
- $\forall i \in [n], X_i$ is the (unordered) score of item i. X_is are *iid* of density f_X.
- $\forall k \in [m], Y_k$ is the k-th smallest threshold, *i.e.* the k-th order statistic of m *iid* random variables of density f_Y (with convention $Y_0 = 0$, and $Y_{k+1} = 1$).
- $\forall k \in [m] \cup \{0\}, B_k \triangleq |\{i \in [n] | X_i \in [Y_k, Y_{k+1}]\}|$, the number of items in bin k.
- $B \triangleq B_K, K \sim \mathcal{U}([m] \cup \{0\})$, the number of items in a random bin, chosen uniformly at random among the bins. We define in the same way $Y \triangleq Y_K$ and $D \triangleq D_K$, a random threshold and the length of a random bin.
- F: the value of the MSF.

f_X and f_Y are assumed to take non-zero values on $(0, 1)$.

In order to simplify the analysis, we define the *rescaled* thresholds and item scores, by composing everything by the *cdf* F_Y. Thanks to this manipulation, the *rescaled* thresholds are uniform on $[0, 1]$, and the item scores remain *iid*.

Formally, we define $\forall i \in [n], \quad X'_i \triangleq F_Y(X_i)$ and $\forall k \in [m], \quad Y'_k \triangleq F_Y(Y_k)$

So,

$$\forall x, y \in [0, 1], \quad f_{X'}(x) = \frac{f_X(F_Y^{-1}(x))}{f_Y(F_Y^{-1}(x))}, \quad f_{Y'}(y) = \frac{f_Y(F_Y^{-1}(y))}{f_Y(F_Y^{-1}(y))} = 1$$

We can define all the corresponding *rescaled* objects:

- $f_{X'}$: the density of the *rescaled* item scores on $[0, 1]$.

- $f_{Y'}$: the density of the *rescaled* user thresholds on $[0,1]$. As proven above, $f_{Y'} = 1$.
- $\forall i \in [n], X'_i \triangleq F_Y(X_i)$ is the (unordered) *rescaled* score of item i. X'_is are *iid* of density $f_{X'}$.
- $\forall k \in [m], Y'_k \triangleq F_Y(Y_k)$ is the k-th smallest *rescaled* threshold .
- $\forall k \in [m] \cup \{0\}, D'_k \triangleq Y'_{k+1} - Y'_k$ the *rescaled* length of bin number k (with $Y'_0 = F_Y(Y_0) = 0$ and $Y'_{m+1} = F_Y(Y_{m+1}) = 1$).
- $\forall k \in [m] \cup \{0\}, B'_k \triangleq |\{i \in [n] | X'_i \in [Y'_k, Y'_{k+1}]\}| = B_k$, the number of items in bin k.
- $B' \triangleq B'_K = B, K \sim \mathcal{U}([m] \cup \{0\})$, the number of items in a random bin, chosen uniformly at random among the bins. We define in the same way $Y' \triangleq Y'_K$ and $D' \triangleq D'_K$, a random *rescaled* threshold and the length of a random *rescaled* bin. Note that we have $Y' = F_Y(Y)$, because $\forall k \in [m], \quad Y'_k \triangleq F_Y(Y_k)$.
- $F' = F$: the value of the MSF.

Because this transformation is increasing, the order of the thresholds and the items is preserved. As a consequence, the MSF remains unchanged. This means that we can reason under the assumption that $f_Y = 1$ without loss of generality, because the analysis of the MSF given any (f_X, f_Y) is equivalent to the analysis given $(f_{X'}, 1)$, as defined above. Under this assumption, our results are expressed in function of $\mathbb{E}[f_X(Y)^2]$, and converted to the full quadratic divergence of the general case using Lemma 26.

G.2. Properties of the quadratic divergence

In the following lemma, we show that the $\mathbb{E}[f_X(Y)^2]$ term becomes the $\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right]$ term of Theorems 5 and 6 for $f_Y \neq 1$.

Lemma 26 *Let (f_X, f_Y) be the true densities of the scores and thresholds, and $(f_{X'}, 1)$ be their rescaled counterparts. Then*

$$\mathbb{E}[f_{X'}(Y')^2] = \mathbb{E}\left[\left(\frac{f_X(Y)}{f_Y(Y)}\right)^2\right]$$

Proof

we have
$\forall i \in [n], X'_i \triangleq F_Y(X_i)$ and $\forall k \in [m], Y'_k \triangleq F_Y(Y_k)$, so
$X' = F_Y(X)$ and $Y' = F_Y(Y)$
So,

$$\forall x \in [0,1], \quad f_{X'}(x) = \frac{f_X(F_Y^{-1}(x))}{f_Y(F_Y^{-1}(x))}$$

$$\Rightarrow f_{X'}(Y') = \frac{f_X(F_Y^{-1}(Y'))}{f_Y(F_Y^{-1}(Y'))} = \frac{f_X(Y)}{f_Y(Y)}$$

$$\Rightarrow \mathbb{E}[f_{X'}(Y')^2] = \mathbb{E}\left[\left(\frac{f_X(Y)}{f_Y(Y)}\right)^2\right]$$

■

As discussed in Section 3.1, the intuition suggests that the ideal case is when the score and item distributions are the same. In the following lemma, we show that the divergence term is indeed minimal *iif* $f_X = f_Y$.

Lemma 27

$$\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right] \geq 1$$

and $\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right] = 1$ iif $f_X(Y) = f_Y(Y)$ a.s..

Proof As explained at the beginning of the Section, we can prove the results with the assumption $f_Y = 1$ without loss of generality. By convexity, we have:

$$\mathbb{E}[f_X(Y)^2] \geq \mathbb{E}[f_X(Y)]^2 = \left(\int_0^1 f_X(y)dy\right)^2 = 1$$

with equality *iif* $f_X = 1$, *i.e.* $f_X = f_Y$. ■

Finally, we look at the particular case where X and Y follow Beta distributions in order to get a closed form for this divergence term.

Lemma 28

Let X and Y be two Beta random variables with $X \sim Beta(a_X, b_X)$ and $Y \sim Beta(a_Y, b_Y)$, where $a_X > \frac{b_Y}{2} > 0$, $a_X \geq \frac{b_X}{2} > 0$. Then we have

$$\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right] = \frac{\mathbf{B}(a_Y, b_Y)}{\mathbf{B}(a_X, b_X)^2}\mathbf{B}(2a_X - a_Y, 2b_X - b_Y)$$

where $\mathbf{B}$ is the Beta function.

Proof

$$\begin{aligned}
\mathbb{E}\left[\frac{f_X(Y)^2}{f_Y(Y)^2}\right] &\triangleq \mathbb{E}_Y\left[\left(\frac{f_X(Y)}{f_Y(Y)}\right)^2\right] \\
&= \int_0^1 \left(\frac{f_X(y)}{f_Y(y)}\right)^2 f_Y(y)dy \\
&= \int_0^1 \frac{f_X(y)^2}{f_Y(y)}dy \\
&= \frac{\mathbf{B}(a_Y, b_Y)}{\mathbf{B}(a_X, b_X)^2}\int_0^1 \frac{y^{2(a_X-1)}(1-y)^{2(b_X-1)}}{y^{a_Y-1}(1-y)^{b_Y-1}}dy \\
&= \frac{\mathbf{B}(a_Y, b_Y)}{\mathbf{B}(a_X, b_X)^2}\int_0^1 y^{2a_X-a_Y-1}(1-y)^{2b_X-2b_Y-1}dy \\
&= \frac{\mathbf{B}(a_Y, b_Y)}{\mathbf{B}(a_X, b_X)^2}\mathbf{B}(2a_X - a_Y, 2b_X - b_Y)
\end{aligned}$$

■

Appendix H. Appendix of Section 4, on the TBS Algorithm

H.1. Full TBS Algorithm

We detail here the three submodules of TBS (Algorithm 2): Algorithms 3, 4 and 5. In the algorithms, we use the notation "Query rating $q(u,i)$" to indicate that the algorithm requests the rating of item i from user u.

Algorithm 2: TBS(n,m)

$\mathcal{B}_1 \leftarrow [n]$; /* bin */
$\mathfrak{B} \leftarrow (\mathcal{B}_1)$; /* bin sequence */
for u *going from* 1 *to* m **do**
  $l, r \leftarrow$ SEARCH$(\mathfrak{B}, u)$; /* find pair of bins containing Y_u (Algo 3) */
  $k^* \leftarrow$ ISOLATE$(\mathfrak{B}, l, r, u)$; /* identify the correct one (Algo 4) */
  $\mathcal{B}^-, \mathcal{B}^+ \leftarrow$ SPLIT$(\mathcal{B}_{k^*}, u)$; /* try to split the bin in two (Algo 5) */
  if $|\mathcal{B}^-| > 0$ *and* $|\mathcal{B}^+| > 0$ **then**
    $\mathfrak{B} \leftarrow (\mathcal{B}_1, ..., \mathcal{B}_{k^*-1}, \mathcal{B}^-, \mathcal{B}^+, \mathcal{B}_{k^*+1}, ..., \mathcal{B}_{|\mathfrak{B}|})$
  end
end
Return $\mathfrak{B}$

Algorithm 3: SEARCH$(\mathfrak{B}, u)$
Find a subset of two adjacent bins of which one contains the threshold.

$l \leftarrow 1$
$r \leftarrow |\mathfrak{B}|$
while $l < r - 1$ **do**
  $k \leftarrow \lfloor \frac{l+r}{2} \rfloor$
  $i \leftarrow$ uniformly sampled item from $\mathcal{B}_k$
  Query rating $q(u,i)$
  if $q(u,i) = 1$ **then**
    $r \leftarrow k$
  else
    $l \leftarrow k$
  end
end
Return l, r

H.2. Detailed Idea of Proof of the Upper Bound on the Complexity of TBS

In this section, we make the distinction between Y_u, the threshold of user u, and $Y_{(k)}$, the k-th smallest user threshold.

Algorithm 4: Find which bin out of $\mathcal{B}_l$ and $\mathcal{B}_r$ is the one containing Y_u.

$R_l \leftarrow \emptyset$
$R_r \leftarrow \emptyset$
while $\mathcal{B}_l$ *and* $\mathcal{B}_r$ *both contain an item that has not been rated by user* u **do**
  $i \leftarrow$ uniformly sampled item from $\mathcal{B}_l \backslash R_l$
  $j \leftarrow$ uniformly sampled item from $\mathcal{B}_r \backslash R_r$
  $R_l \leftarrow R_l \cup \{i\}$
  $R_r \leftarrow R_r \cup \{r\}$
  `Query` ratings $q(u,i)$ and $q(u,j)$
  if $q(u,i) = 1$ **then**
    Return l
  end
  if $q(u,j) = 0$ **then**
    Return r
  end
end
if $|\mathcal{B}_l| > |\mathcal{B}_r|$ **then**
  Return l
else
  Return r
end

Algorithm 5: `SPLIT`$(\mathcal{B}, u)$
Split the selected bin.

for i *in* $\mathcal{B}$ **do**
  `Query` rating $q(u,i)$
end
$\mathcal{B}^- \leftarrow \{i \in \mathcal{B} \mid q(u,i) = 0\}$
$\mathcal{B}^+ \leftarrow \{i \in \mathcal{B} \mid q(u,i) = 1\}$
Return $\mathcal{B}^-, \mathcal{B}^+$

We recall that, as explained in Section 4, we split the total query cost Q in the following way:

$$Q = Q^{search} + Q^{iso} + Q^{split} \tag{31}$$

where Q^{search}, Q^{iso}, and Q^{split} are the number of queries performed respectively during the SEARCH, ISOLATE and SPLIT phases.

In the specification of Algorithm 2, nothing prevents the algorithm from making several times the same query (in different phases). However, it is reasonable to consider that the ratings given by the users are stored, and that asking a second time the same rating to the same user does not count as a query.

Under this assumption, we define Q^{iso} and Q^{split} such that we count all the queries on the items of bin $B_{k^*(u)}$ in Q^{split} and not in Q^{iso}, where $B_{k^*(u)}$ is the bin being split at step u. Note that this change of definition does not affect equation (31). In what follows, we refer by $\texttt{SEARCH}_u$, $\texttt{ISOLATE}_u$, $\texttt{SPLIT}_u$ and $\texttt{UPDATE}_u$ to the execution of these phases at step u.

We will now upper bound the expectations of these three random variables. For this, we use the three following lemmas:

Lemma 29 $\mathbb{E}[Q^{search}] \leq m(\log_2(n) + 1)$

Lemma 30 $Q^{iso} \leq Q^{split}$

Lemma 31 $\mathbb{E}[Q^{split}] \lesssim 2n\log(m) + 2m$

We restate and prove each of these three lemmas:

Lemma 29 $\mathbb{E}[Q^{search}] \leq m(\log_2(n) + 1)$

Proof For each user, we perform a binary search on the current set of non-empty bins. There cannot be more non-empty bins than the number of items n. Binary search over n bins takes at most $\lceil \log_2(n) \rceil$ queries. Therefore, each user can do at most $\lceil \log_2(n) \rceil$ ratings during the SEARCH phase of user u.

This gives us:

$$\begin{aligned} Q^{search} &\leq \sum_{u=1}^{m} \lceil \log_2(n) \rceil \\ &\leq m(\log_2(n) + 1) \end{aligned}$$

and in particular

$$\mathbb{E}[Q^{search}] \leq m(\log_2(n) + 1)$$

■

Lemma 30 $Q^{iso} \leq Q^{split}$

Proof For a given user u, let Q_u^{iso} (*resp.* Q_u^{split}) be the number of ratings performed by u during `ISOLATE` (*resp.* `SPLIT`).

Suppose $\mathcal{B}_r$ is returned by $\texttt{ISOLATE}_u$. In reality, in $\texttt{ISOLATE}_u$, we rate some (or all) items in $\mathcal{B}_l$ and some in $\mathcal{B}_r$. However, as stated at the beginning of the section, for accounting purposes, we count the ratings of $\mathcal{B}_r$ made during `ISOLATE` as part of `SPLIT` (so we bound them in Lemma 31). In `SPLIT`, we anyway rate the rest of $\mathcal{B}_r$. So, we simply say that we rate (some or all) items of $\mathcal{B}_l$ in $\texttt{ISOLATE}_u$ and all of $\mathcal{B}_r$ in $\texttt{SPLIT}_u$.

Now there are two cases: either $\texttt{ISOLATE}_u$ stopped after rating all items from $\mathcal{B}_l$ or it stopped before.

In the first case, the construction of `ISOLATE` implies that $|\mathcal{B}_l| \leq |\mathcal{B}_r|$. By assumption, we have $Q_u^{split} = |\mathcal{B}_r|$.
This gives $Q_u^{iso} \leq |\mathcal{B}_l| \leq |\mathcal{B}_r| = Q_u^{split}$.

In the second case, the construction of `ISOLATE` implies that we rated an equal number of items of $\mathcal{B}_l$ and $\mathcal{B}_r$ during `ISOLATE`. However, the ones of $\mathcal{B}_r$ are counted in Q_u^{split}. This gives us $Q_u^{iso} \leq Q_u^{split}$.

Consequently, in all cases, we have $Q_u^{iso} \leq Q_u^{split}$.
As this is true for every u, this directly yields $Q^{iso} = \sum_{u=1}^{m} Q_u^{iso} \leq \sum_{u=1}^{m} Q_u^{split} = Q^{split}$. ■

Lemma 31 $\mathbb{E}[Q^{split}] \lesssim 2n\log(m) + 2m$

Proof [Idea of proof]

We use some inexact approximations to prove the inequality (hence the $\lesssim$ in the statement).

At a given step u of the algorithm (after user $u-1$), for a given threshold index k we define:

- $D_k(u) \triangleq Y_{(k+1)}(u) - Y_{(k)}(u)$ the *length* of bin k, *i.e.* the distance between two consecutive thresholds.
- $X_k^+(u)$, the smallest item score bigger than $Y_{(k+1)}(u)$.
- $X_k^-(u)$, the biggest item score smaller than $Y_{(k)}(u)$.
- $D_k^+(u) \triangleq X_k^+(u) - X_k^-(u)$.
- $d_k^-(u) \triangleq Y_{(k)}(u) - X_k^-(u)$.
- $d_k^+(u) \triangleq X_k^+(u) - Y_{(k+1)}(u)$.

This new notation is illustrated on Figure 3.

Let $Y(u) = (Y_{(1)}(u), \ldots, Y_{(u-1)}(u))$ be the sequence of thresholds used by the algorithm after step $u-1$, and X the sequence of item scores.

Note that $D_k(u) \sim \text{Beta}(1, u-1)$, because it is the difference between two consecutive order statistics of a $\mathcal{U}([0,1])$, cf Lemma 22 (proven below in the Appendix).

Let $k^*(u)$ the index of the bin returned by $\texttt{ISOLATE}_u$. A bin $\mathcal{B}_k$ can be selected by $\texttt{ISOLATE}_u$ only if the threshold of user u belongs to $[X_k^-(u), X_k^+(u)]$. So, for all k, u, we have the inclusion:

$$(k^*(u) = k) \subseteq (Y_u \in [X_k^-, X_k^+]) \tag{32}$$

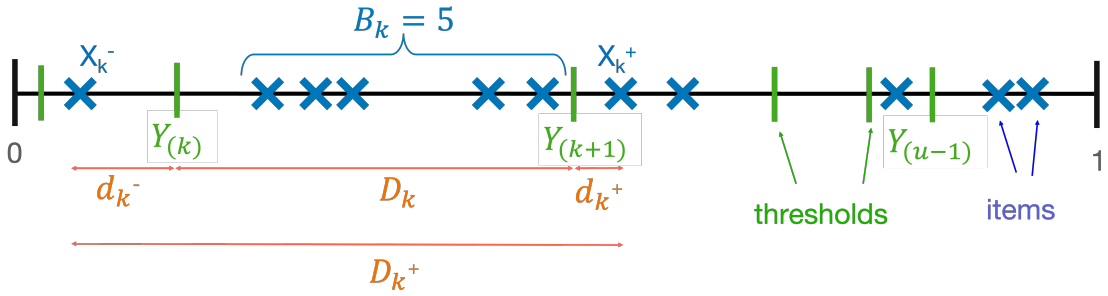


Figure 3: Illustration of the definitions for u fixed. We omit the dependency in u on the figure for readability. All random variables here depend on u.

We decompose the expected cost in queries of the SPLIT phase (Q^{split}) by the expected cost for each step u of the algorithm (Q_u^{split}). For each step, we further decompose Q_u^{split} by the expected cost for each bin of this step, indexed by k.
Note that we consider all pair of consecutive thresholds as a bin, even if the bin contains no item. Recall that $\mathcal{B}_k(u)$ is the *size* of the bin (*i.e.* the number of item scores between $Y_{(k)}(u)$ and $Y_{(k+1)}(u)$). Then we have:

$$\mathbb{E}[Q_u^{split} \mid Y(u)] = \sum_{k=0}^{u} \mathbb{P}(k^*(u) = k \mid Y(u)) \mathbb{E}[B_k(u) \mid k^*(u) = k, Y(u)] \tag{33}$$

$$\leq \sum_{k=0}^{u} \mathbb{P}(Y_u \in [X_k^-, X_k^+] \mid Y(u)) \mathbb{E}[B_k(u) \mid Y_u \in [X_k^-, X_k^+], Y(u)] \tag{34}$$

$$= \sum_{k=0}^{u} \mathbb{E}[D_k^+(u) \mid Y(u)] \mathbb{E}[B_k(u) \mid Y(u)] \tag{35}$$

$$= \sum_{k=0}^{u} \mathbb{E}[D_k^+(u) \mid Y(u)] n D_k(u) \tag{36}$$

$$= n \sum_{k=0}^{u} \mathbb{E}[D_k^+(u) \mid Y(u)] D_k(u) \tag{37}$$

where (34) uses Lemma 32 (proven below in the Appendix) and equation (32).

By definition, $D_k^+(u) = D_k + d^- + d^+$. So

$$\begin{aligned}\mathbb{E}[D_k^+(u) \mid Y(u)] D_k(u) &= \mathbb{E}[D_k(u) + d_k^+ + d_k^- \mid Y(u)] D_k(u) \\ &= D_k(u)^2 + \mathbb{E}[d_k^+ + d_k^- \mid Y(u)] D_k(u)\end{aligned}$$

So,

$$\begin{aligned}\mathbb{E}[Q_u^{split}] &= \mathbb{E}[\mathbb{E}[Q_u^{split} \mid Y(u)]] \\ &\leq \mathbb{E}\left[n\sum_{k=0}^{u}(D_k(u)^2 + \mathbb{E}[d_k^+ + d_k^- \mid Y(u)]D_k(u))\right] \\ &= \mathbb{E}\left[n(u+1)(D_K(u)^2 + \mathbb{E}[d_K^+ + d_K^- \mid Y(u)]D_K(u))\right]\end{aligned}$$

where K is uniform on $[u] \cup \{0\}$ and independent of the thresholds and items.

Conditional on set of thresholds $Y(u)$, regardless of their positions in $[0,1]$, the expected distance between any given threshold and its closest item score on the right (or the left) is roughly $\frac{1}{n+1}$. The approximation is even more valid as the number of users m grows, because K is the index of one of these thresholds selected uniformly at random, so d_K^+ is approximately the distance between two consecutive points selected uniformly and independently at random in $[0,1]$.

So we have $\mathbb{E}[d_K^+ + d_K^- \mid Y(u)] \simeq \frac{2}{n+1}$.

$$\begin{aligned}\mathbb{E}[Q_u^{split}] &\leq n(u+1)\left(\mathbb{E}[D_K(u)^2] + \mathbb{E}[\mathbb{E}[d_K^+ + d_K^- \mid Y(u)]D_K(u))]\right) \\ &\simeq n(u+1)\left(\mathbb{E}[D_K^2] + \frac{2}{n+1}\mathbb{E}[D_K]\right) \\ &= n(u+1)\left(\frac{2}{(u+1)(u+2)} + \frac{2}{n+2}\frac{1}{u+1}\right) \\ &= 2\frac{n}{u+2} + 2\frac{n}{n+2}\end{aligned}$$

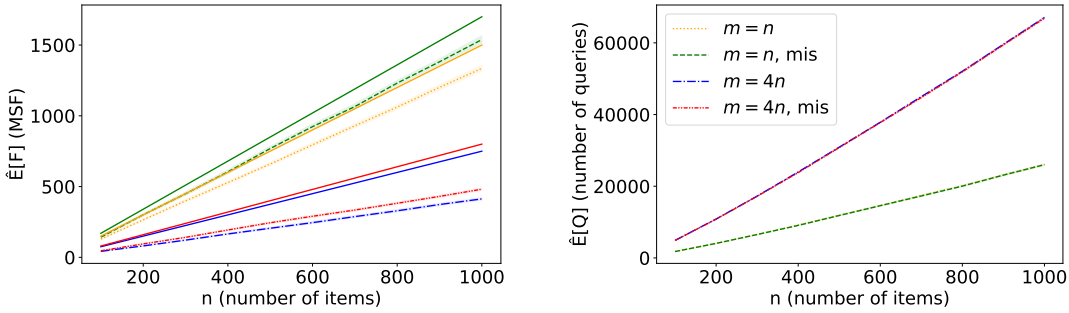
We sum on all users to get the final result:

$$\begin{aligned}\mathbb{E}[Q^{split}] &= \sum_{u=1}^{m}\mathbb{E}[Q_u^{split}] \\ &\lesssim \sum_{u=1}^{m}\left(2\frac{n}{u+2} + 2\frac{n}{n+2}\right) \\ &\simeq 2n\log(m) + 2m\end{aligned}$$

■

Appendix I. Additional Experiments

Linear regime We can see on Figure 4(*a*) that the expected MSF is linear in the number of items, as stated by Theorem 5. Additionally, we observe that the MSF is smaller than the estimation of $n(\frac{1}{2} + \frac{1}{r}(f_X(Y)/f_Y(Y))^2)$ of Theorem 5, with the gap growing linearly. Theorem 5 showed that we have $\mathbb{E}[F] = n(\frac{1}{2} + \frac{1}{r}(f_X(Y)/f_Y(Y))^2) + g(n)n + o(n)$ with $|g| \leq \frac{r}{2}$. According to Figure



(*a*) Expected MSF (*b*) Expected number of queries

Figure 4: Experiments on the expected MSF and number of queries for a number of users linear in the number of items. The full lines on Figure (a) are the theoretical values of Theorem 5. Figure (b) shows the empirical query cost of TBS. The label *'mis'* indicates the experiments with a mismatch between the distributions of the items and the thresholds. We use $a_X = 2, b_X = 3, a_Y = 2, b_Y = 2$ for the mismatch case, and $a_X = 1, b_X = 1, a_Y = 1, b_Y = 1$ in the default case. On Figure (b), confidence intervals are too small to display, and lines for the mismatch case overlap with their counterparts for the matching case.

4(*a*) we most likely have $g < 0$. We can see on Figure 4(*b*) that the number of queries of TBS follows the $m \log(n)$ tendency discussed in Section 4 ($n \log(n)$ here), which confirms the analysis of Lemmas 29, 30 and 31. We also observe that the mismatch in the linear regime has very little effect on the number of queries per user. This indicates that even though our upper bound analysis is limited to equal distributions for the items and thresholds, the approximate upper bound we provide is still reasonable in the mismatched case.

Appendix J. Inequalities

We regroup here a few general inequalities which are used in the proofs of other lemmas.

Lemma 32 *Let $A \subseteq B$ be two probabilistic events and X a positive random variable, then*

$$\mathbb{P}(A)\mathbb{E}[X \mid A] \leq \mathbb{P}(B)\mathbb{E}[X \mid B]$$

Proof We have $A \subseteq B$, so $B = A \cup (B \backslash A)$. So:

$$\begin{aligned}
\mathbb{P}(B)\mathbb{E}[X|B] &= \mathbb{P}(B)\mathbb{E}[X|A \cup (B\backslash A)] \\
&= \mathbb{P}(B)(\mathbb{P}(A|B)\mathbb{E}[X|A] + \mathbb{P}(B\backslash A|B)\mathbb{E}[X|B\backslash A]) && \text{(because } A \text{ and } B\backslash A \text{ are disjoint)} \\
&\geq \mathbb{P}(B)\mathbb{P}(A|B)\mathbb{E}[X|A] && \text{(because } X \geq 0\text{)} \\
&= \mathbb{P}(A)\mathbb{E}[X|A]
\end{aligned}$$

■

Lemma 33

For all $k \leq n \leq m$

$$\frac{\binom{n}{k}}{\binom{k+m}{k}} \leq \frac{e^2}{2\pi}\left(\frac{n}{m}\right)^k$$

Proof

For the case $k = n \leq m$, we have

$$\left(\frac{\binom{n}{k}}{\binom{k+m}{k}}\right)^{-1} = \binom{n+m}{n} = \prod_{j=1}^{n}\frac{m+j}{j} = \prod_{j=1}^{n}\left(1+\frac{m}{j}\right) \geq \prod_{j=1}^{n}\left(1+\frac{m}{n}\right) \geq \left(\frac{m}{n}\right)^n \geq \frac{2\pi}{e^2}\left(\frac{m}{n}\right)^k$$

For remaining cases $k < n \leq m$, we use the bounds $\sqrt{2\pi}n^{n+1/2}e^{-n} \leq n! \leq en^{n+1/2}e^{-n}$. They give us

$$\begin{aligned}
\frac{\binom{n}{k}}{\binom{k+m}{k}} &= \frac{n!}{(n-k)!}\frac{m!}{(k+m)!} \\
&\leq \frac{en^{n+1/2}e^{-n}}{\sqrt{2\pi}(n-k)^{n-k+1/2}e^{-(n-k)}}\frac{em^{m+1/2}e^{-m}}{\sqrt{2\pi}(k+m)^{k+m+1/2}e^{-(k+m)}} \\
&\leq \frac{e^2}{2\pi}\frac{n^{n+1/2}}{(n-k)^{n-k+1/2}}\frac{m^{m+1/2}}{(k+m)^{k+m+1/2}} \\
&= \frac{e^2}{2\pi}\left(\frac{n-k}{n}\right)^{-n}\left(\frac{m+k}{m}\right)^{-m} n^{1/2}(n-k)^{k-1/2}m^{1/2}(m+k)^{-k-1/2} \\
&\leq \frac{e^2}{2\pi}\left(1-\frac{k}{n}\right)^{-n}\left(1+\frac{k}{m}\right)^{-m}\left(\frac{n-k}{m+k}\right)^k \\
&= \frac{e^2}{2\pi}\left(\frac{n}{m}\right)^k\left(1-\frac{k}{n}\right)^{k-n}\left(1+\frac{k}{m}\right)^{-k-m}
\end{aligned}$$

We just need to show that $\left(1-\frac{k}{n}\right)^{k-n}\left(1+\frac{k}{m}\right)^{-k-m} \leq 1$. For this, we will show that its logarithm is negative. We have

$$\log\left(\left(1-\frac{k}{n}\right)^{k-n}\left(1+\frac{k}{m}\right)^{-k-m}\right) = (k-n)\log\left(1-\frac{k}{n}\right) - (k+m)\log\left(1+\frac{k}{m}\right)$$

In addition, for $0 < x \leq 1$, we have $-\log(1-x) \leq x + \frac{x^2}{2(1-x)}$ and $\log(1+x) \geq x - \frac{x^2}{2}$, which respectively give us

$$(k-n)\log(1-\frac{k}{n}) \leq (n-k)\left(\frac{k}{n} + \frac{(\frac{k}{n})^2}{2(1-\frac{k}{n})}\right) = k - \frac{k^2}{2n}$$

$$-(k+m)\log(1+\frac{k}{m}) \leq -(m+k)\left(\frac{k}{m} - \frac{(\frac{k}{m})^2}{2}\right) = -k - \frac{k^2}{2m} + \frac{k^3}{2m^2}$$

Summing the two terms, we obtain

$$(k-n)\log(1-\frac{k}{n}) - (k+m)\log(1+\frac{k}{m}) \leq -\frac{k^2}{2n} - \frac{k^2}{2m} + \frac{k^3}{2m^2}$$

For n large enough, we have $k \leq n \leq m$, which gives $-\frac{k^2}{2n} + \frac{k^3}{2m^2} \leq 0$, so finally

$$(k-n)\log\left(1-\frac{k}{n}\right) - (k+m)\log\left(1+\frac{k}{m}\right) \leq -\frac{k^2}{2m} < 0$$

■

Lemma 34 *Let Z be a random variable and $\mathcal{A}$ be a probabilistic event. Then*

$$|\mathbb{E}[Z] - \mathbb{E}[Z|\mathcal{A}]| \leq (1-\mathbb{P}(\mathcal{A}))|\mathbb{E}[Z|\bar{\mathcal{A}}] - \mathbb{E}[Z|\mathcal{A}]|$$

in particular, if $0 \leq Z \leq c_Z$, then

$$|\mathbb{E}[Z] - \mathbb{E}[Z|\mathcal{A}]| \leq (1-\mathbb{P}(\mathcal{A}))c_Z$$

Proof

$$\mathbb{E}[Z] = \mathbb{P}(\mathcal{A})\mathbb{E}[Z|\mathcal{A}] + (1-\mathbb{P}(\mathcal{A}))\mathbb{E}[Z|\bar{\mathcal{A}}]$$

so

$$\begin{aligned}|\mathbb{E}[Z] - \mathbb{E}[Z|\mathcal{A}]| &= |(\mathbb{P}(\mathcal{A})-1)\mathbb{E}[Z|\mathcal{A}] + (1-\mathbb{P}(\mathcal{A}))\mathbb{E}[Z|\bar{\mathcal{A}}]| \\ &= (1-\mathbb{P}(\mathcal{A}))|\mathbb{E}[Z|\mathcal{A}] - \mathbb{E}[Z|\bar{\mathcal{A}}]|\end{aligned}$$

If $0 \leq Z \leq c_Z$, then $|\mathbb{E}[Z] - \mathbb{E}[Z|\mathcal{A}]| \leq c_Z$, so

$$|\mathbb{E}[Z] - \mathbb{E}[Z|\mathcal{A}]| = (1-\mathbb{P}(\mathcal{A}))c_Z$$

■