Discrete Choice, Permutations, and Reconstruction

Flavio Chierichetti*
Sapienza University of Rome
Rome, Italy
flavio@di.uniroma1.it

Ravi Kumar
Google
Mountain View, CA
ravi.k53@gmail.com

Andrew Tomkins
Google
Mountain View, CA
atomkins@gmail.com

Abstract

In this paper we study the well-known family of *Random Utility Models*, developed over 50 years ago to codify rational user behavior in choosing one item from a finite set of options. In this setting each user draws i.i.d. from some distribution a utility function mapping each item in the universe to a real-valued utility. The user is then offered a subset of the items, and selects the one of maximum utility. A Max-Dist oracle for this choice model takes any subset of items and returns the probability (over the distribution of utility functions) that each will be selected. A discrete choice algorithm, given access to a Max-Dist oracle, must return a function that approximates the oracle.

We show three primary results. First, we show that any algorithm exactly reproducing the oracle must make exponentially many queries. Second, we show an equivalent representation of the distribution over utility functions, based on permutations, and show that if this distribution has support size k, then it is possible to approximate the oracle using $O(nk)$ queries. Finally, we consider settings in which the subset of items is always small. We give an algorithm that makes less than $n^{(1-\epsilon/2)K}$ queries, each to sets of size at most $(1-\epsilon/2)K$, in order to approximate the Max-Dist oracle on every set of size $|T| \leq K$ with statistical error at most ϵ. In contrast, we show that any algorithm that queries for subsets of size $2^{O(\sqrt{\log n})}$ must make maximal statistical error on some large sets.

1 Introduction

In this paper we study the problem of *discrete choice*, in which a user must select exactly one element from a discrete set of alternatives. Discrete choice models are widely used in marketing (choice of a brand), urban planning (choice of transportation), politics (choice of a leader in an election), finance (choice of investments), and many other fields. Nonetheless, little is know about the availability of algorithms with provable guarantees, or the hardness of answering or approximating basic questions such as the likelihood of a particular choice.

We begin with a brief introduction to the classical setting, and then give an overview of our results. Let $[n]$ be the universe of items and $\varnothing \neq T \subseteq [n]$ be the *choice set* of options available to a decision maker (also for brevity called a *user*). A *discrete choice model* is a function f mapping from a choice set T to a distribution over the elements of T indicating the probability that each element will be chosen. In general, for subsets $S, T \subseteq [n]$, the distributions $f(S)$ and $f(T)$ need have no connection, even if S and T are highly overlapping. Information on $f(T)$ for all subsets $T \neq S$ will give no information about $f(S)$. In real-life applications, however, there are typically exploitable connections between a user's behavior when faced with a choice set T versus a highly overlapping choice set T'.

Random utility models. There is a large and highly-studied subclass of discrete choice models called *Random Utility Models*, or RUMs, that impose some lightweight consistency constraints on overlapping choice sets. In these models, each successive choice is undertaken by a new user, drawn i.i.d. from a distribution of users. The user is represented by a vector $\vec{u} \in \mathbb{R}^n$ representing the utility this user will accrue by selecting each item of the universe. For a choice set T, the user behaves rationally by selecting $\arg\max_{t \in T} u[t]$. For each successive choice, a new utility vector is drawn i.i.d. from the joint distribution, a new choice set is presented, and a new decision is registered. This model dates back to work of Marschak in 1960 [10], and is well studied in the behavioral sciences community.

Two overlapping subsets T and $T' = T \cup \{i\}$ may be presented to users with different utility functions, yielding very different choice patterns, but the fact that utility functions are drawn independently from the *same* distribution suggests that the distributions for T and T' are now connected in some way. For example, by construction, it is clear that for any element $j \in T$,

*Work done in part while visiting Google. Supported in part by a Google Focused Research Award, by the ERC Starting Grant DMAP 680153, and by the SIR Grant RBSI14Q743.

we must have $\Pr[j \mid T] \geq \Pr[j \mid T']$. Similarly, $\Pr[j \mid T] - \Pr[j \mid T'] \leq \Pr[i \mid T']$. Such observations suggest that RUMs are much more constrained than general discrete choice models, and hence raise the question of learning a RUM from queries.

(For readers familiar with discrete choice, we should provide one paragraph of additional context. Classically, the utility $u[t]$ is usually broken into two components: $u[t] = v_t + \varepsilon_t$. v_t is based on factors known to the algorithm, while ε_t is a random noise variable representing the error in the algorithm's knowledge. v_t is sometimes taken to be a linear combination of known features of the current context, or is sometimes taken to be a constant underlying quality of item t, fixed across all the choices we observe. We adopt this latter view, which justifies our decision to treat the vector of utilities $\vec{u}$ as a random variable drawn i.i.d. for each choice from a joint utility distribution.)

As the user is rational, two utility vectors that order the elements of $[n]$ identically will yield the same decisions for every choice set. Thus, any RUM may be presented as a distribution over the set of permutations of $[n]$, rather than over the continuous space $\mathbb{R}^n$; we adopt the former more manageable representation henceforth. Our goal is to study whether it is possible, given access to one of several possible oracles for a particular RUM, to return a function mapping choice sets to distributions of the probability each item will be selected.

1.1 Our contributions

In this paper we make the following contributions. First we show that $\Omega(2^n)$ queries are required to exactly reconstruct a RUM. Next, we show that RUMs representable using only k permutations may be learned in $O(nk)$ queries. We then show that queries to choice sets of bounded size yield maximally bad approximations in total variation distance on large sets. However, with knowledge of the distribution for all choice sets of size k, an algorithm may approximate distributions for choice sets of size up to K with total variation distance at most $1 - k/K$. Finally, we show that any RUM may be approximated to any constant total variation distance by another RUM with $O(n^2)$ permutations; we discuss below why this does not contradict our first two contributions above. We now give a brief overview of these results.

Exact reconstruction lower bound. We begin in Section 3 with a general lower bound showing that, for an arbitrary RUM, exponentially many choice sets must be observed in order to reconstruct for every choice set the distribution of each item being selected within a total variation distance of $o(2^{-(3/2)n})$. This negative result implies that it is not possible to develop a general exact algorithm for determining choice probabilities in RUMs.

Algorithm for bounded support. Our main positive result gives an algorithm (Section 4) to learn RUMs with support over only polynomially many permutations. The result is as follows: given a RUM that is expressible as a distribution over k permutations, there is an algorithm that makes $O(nk)$ queries to an oracle providing the exact likelihood that each element of a choice set will be chosen, and returns a function that will provide the exact choice distribution for every choice set. This algorithm does not directly represent the permutations; in fact, we show that there exist RUMs that can be represented using any of two disjoint distributions over permutations.

Small subsets bounds. In common practical settings, users are offered choice sets T that are small relative to the entire universe $[n]$, e.g., a video recommender system that may offer 100 movies to the user from a database of millions. In this context, we give an algorithm (Section 5.2) such that, for any K and any $\epsilon \in [0, 1]$, with less than $n^{(1-\epsilon)K}$ queries, each to sets of size at most $(1 - \epsilon)K$, we can determine the choice distribution of every $|T| \leq K$ with total variation error ϵ. In contrast, we show (Section 5.1) that based on arbitrarily many examples of choices over sets T with $|T| \leq \lg n$, any algorithm must make maximal total variation error on some (large) sets.

Concise representations. We show in Section 6 that any RUM may be approximated with total variation ϵ for any $\epsilon > 0$ by another RUM supported on only $O(n^2/\epsilon^2)$ permutations. Using this coreset-type result, we show that polynomial-size *mixed logit* models may approximate any RUM, which improves over the exponential-size upper bound of McFadden and Train [11].

The coreset result seemingly suggests an algorithm to learn general RUMs by learning its small-support representation instead using our algorithm in Section 4. Unfortunately we show that this approach does not work: using our algorithm to learn the small-support RUM in this fashion would require sampling *directly* from the small-support RUM. As we may only sample from the large-support RUM at total variation distance ϵ, the small errors in approximation would cascade.

Our work may be viewed as a very preliminary step towards establishing a clean theoretical footing for discrete choice algorithms. There remain large gaps between the upper bounds and lower bounds we show, suggesting there is much more to be done in resolving the complexity of discrete choice problems.

1.2 Related work For a small but crucial subclass of RUMs, there is a well-known efficient algorithm that is in common use, called the *multinomial logit*, or MNL. To describe the model, we revisit our earlier observation that the utility $u[t]$ of selecting item t may be decomposed as $u[t] = v_t + \varepsilon_t$. In MNL, the probability that an element t is selected from a choice set is proportional to v_t:

$$\Pr[\text{decision maker selects } t \in T] = \frac{v_t}{\sum_{s \in T} v_s}. \tag{1.1}$$

For two items s and t, (1.1) implies that $\Pr[u[s] > u[t]] = \Pr[v_s + \varepsilon_s > v_t + \varepsilon_t] = \Pr[\varepsilon_s - \varepsilon_t > v_t - v_s] = v_s/(v_s + v_t)$. This condition on the difference of two random variables holds if and only if the noise term ε_t is drawn i.i.d. from a Gumbel distribution with CDF $F(x) = e^{-e^{(-x)}}$; see [13] for a proof.

If we believe that our estimate of utility has error with this doubly-exponential CDF (a belief that may be verified by a number of statistical tests), then MNL may be employed to learn the base utilities v_t by an efficient convex optimization. However, the special structure imposes many restrictions. For example, as observed by Luce in 1959 [8], the inclusion of an additional element in the choice set T must from (1.1) reduce the likelihood of all other elements by the same fraction. In many settings this is unlikely. Consider for example a user choosing between a Vegan restaurant and a Steak house. If another, better quality Steak house is added to the choice set, users who would originally have chosen Steak will probably switch, while users who would originally have chosen Vegan will probably not switch; thus the new Steak house would "cannibalize" probability more heavily from the original Steak house, and multinomial logit would not be appropriate.

As the restrictions on multinomial logit are so strong, the discrete choice community has for the last five decades introduced a wide range of additional models within the RUM framework. Other than MNL, almost all such models are non-convex, do not have algorithms with any guarantees, and are typically solved using simulation. See [13] for a survey.

It is known [11] that so-called *mixed logit* models, which are linear combinations of multiple MNL models, may ϵ-approximate any RUM. However, less is known about provably learning these models and existing heuristics typically require expensive techniques based on simulation for a mixture of small constant number of MNLs,[1] while the best-known approximation required exponentially many mixture components. As a direct corollary of our results, we show that mixed logits with polynomially many MNLs are sufficient to ϵ-approximate any RUM, but the practical difficulty of learning mixed logits of this size remains.

We show that exact learning of general RUMs is impossible with sub-exponential oracle queries. Our main algorithm however is incomparable with the algorithms for MNL, as MNL RUMs in general have exponential support when represented as permutations, while RUMs with polynomial support are not in general MNLs. Likewise, our bounds in Section 6 show that results from sets of one size may be parlayed into results for sets that are a constant factor larger, with some loss in accuracy, but that sets that are significantly larger may have arbitrarily large error.

As discussed before, there has been little work on algorithmic questions in discrete choice. Learning the structure of a nested logit model was studied in [2]. For the problem of learning MNL mixtures, there has been some attempts at algorithms with provable guarantees: the setting of low-dimensional structure was considered in [7], the case when each MNL is "geometric" was solved in [1], and using pairwise comparisons to learn was studied in [12]. Blanchett et al. [3] proposed a choice model based on Markov chains and obtain some algorithmic results for learning in this model; see also [6]. Farias et al. [5] study the problem of learning the "sparsest" RUM, in the sense of fewest permutations, which is consistent with a set of observations. However, none of these is directly related to the questions we address.

2 Preliminaries

2.1 Notation Let $[n] = \{1, \ldots, n\}$ and let $\mathbf{S}_n$ be the set of permutations of the set $[n]$. For a given permutation $\pi \in \mathbf{S}_n$ and for $i \in [n]$, we let $\pi(i) \in [n]$ be the *position* (or *rank*) of element i in π. E.g., if $\pi = (3, 1, 4, 2)$, then $\pi(3) = 1, \pi(1) = 2, \pi(4) = 3$ and $\pi(2) = 4$. We use $2^{[n]}$ to denote the powerset of $[n]$ and $\binom{[n]}{k}$ to denote the set of subsets of $[n]$ of size k.

Given a permutation $\pi \in \mathbf{S}_n$ and $T \subseteq [n]$, let

$$\pi^*(T) = \arg\max_{i \in T} \pi(i),$$

i.e., the maximum element in the subset T according to π.

Let $\mathrm{supp}(D)$ denote the support of a distribution D. We use $x \sim D$ to denote that $x \in \mathrm{supp}(D)$ is sampled according to D. The *total variation distance* between discrete distributions D and D' with $\mathrm{supp}(D) = \mathrm{supp}(D')$ is equal to $|D - D'|_{\mathrm{tv}} = \frac{1}{2} \sum_{x \in \mathrm{supp}(D)} |D(x) - D'(x)| = \frac{1}{2}|D - D'|_1$.

[1] For full disclosure, as we describe below, we have another submission at this conference with a result showing that mixed logit models of exactly two MNL components may be learned exactly in polynomial time.

2.2 Query models Let D be a distribution over $\mathbf{S}_n$. Given a set $\varnothing \subset T \subseteq [n]$, we use D_T to denote the distribution of the random variable $\pi^*(T)$ for $\pi \sim D$. Note that $\mathrm{supp}(D_T) \subseteq T$. A MAX-SAMPLE *oracle* for D is a random function $2^{[n]} \mapsto [n]$ that on an input subset $\varnothing \subset T \subseteq [n]$, returns $x \sim D_T$, i.e., a random element $x \in T$ according to D_T. A MAX-DIST *oracle* for D is a function $2^{[n]} \mapsto \mathbb{R}^n$ that on an input $\varnothing \subset T \subseteq [n]$, returns the entire distribution D_T.

It is clear that the MAX-SAMPLE oracle is practically more meaningful but strictly weaker than the MAX-DIST oracle. However, it is easy to see that we can use the MAX-SAMPLE oracle to approximate the output of the MAX-DIST oracle. Indeed, consider the following definition: an (ϵ, δ)-*approx*-MAX-DIST *oracle* for D is a function that on an input $\varnothing \subset T \subseteq [n]$, returns a distribution[2] $\tilde{D}_T$ such that:

- $\Pr\left[|\tilde{D}_T - D_T|_{\mathrm{tv}} \leq \epsilon\right] \geq 1 - \delta$, and
- $\mathrm{supp}(\tilde{D}_T) \subseteq \mathrm{supp}(D_T)$.

Using sampling and standard tail bounds, the following is immediate.

OBSERVATION 2.1. *A sample of an (ϵ, δ)-approx-MAX-DIST oracle for D, for a given set T, can be obtained with $O(\epsilon^{-2}|T| \log \frac{|T|}{\delta})$ independent calls to a MAX-SAMPLE oracle for D, with T as its input.*

Recall the goal in this work. We are given access to a MAX-SAMPLE oracle (or to a MAX-DIST oracle, or to an approx-MAX-DIST oracle). Our goal to approximately reconstruct D_T for each $T \subseteq [n]$ or for each $T \subseteq [n]$ of a certain size. The computational question then is how to do this with as few oracle accesses as possible, i.e., can we somehow use the values of D_T for several T's to reconstruct D_S for an arbitrary S? As indicated in the Introduction, motivated by practical considerations, we also study settings when we place certain restrictions on the type of oracle accesses—bounds on the size of the input subset to the oracle.

2.3 Relationship to RUMs We make the following observation, which we already stated in the Introduction, explicit here. Recall the definition of RUMs: each choice is associated with a vector of utilities drawn from some joint utility distribution and for a choice set, the decision maker behaves rationally by selecting the element in the choice set with the maximum utility.

OBSERVATION 2.2. *For each RUM there exists a distribution on $\mathbf{S}_n$ inducing the same D_T, for each $\varnothing \subset T \subseteq [n]$.*

[2]Observe that the distribution $\tilde{D}(T)$ is a random variable itself.

Proof. Suppose that a random sample from the RUM produces the utilities $\vec{u}$ and given a subset T, outputs the element $x \in T$ such that $u[x]$ is maximum. The utilities $\vec{u}$ induce a natural total preorder on $[n]$ and let π be a uniform permutation between its linear extensions. Then an equivalent permutation oracle can return $\pi^*(T)$. ■

3 A general lower bound

We first show that, even with the powerful MAX-DIST oracle, one needs to query $\Omega(2^n)$ sets to be able to *exactly* reconstruct all distributions D_T for all $T \subseteq [n]$, with constant probability. Obtaining a similar lower bound in the less powerful MAX-SAMPLE oracle case is much easier; we omit the details in this version.

Our proof proceeds by showing that there exist RUMs that induce the uniform distribution over all subsets except for a special planted subset S of size k, for which a specific element has probability that is slightly larger (by $1/\beta_{n,k}$, defined below) than its neighbors. We start with a technical lemma giving the construction. For a non-empty set T, let U_T be the uniform distribution on set T. Let

$$\beta_{n,k} = \binom{n}{n-k, \lfloor \frac{k-1}{2} \rfloor, \lceil \frac{k-1}{2} \rceil, 1},$$

denote the multinomial coefficient. If $k = cn$, it holds that $1/\beta_{n,k} = \Theta\left(2^{-(c+\mathrm{H}(c))\cdot n}\right)$.

LEMMA 3.1. *Let $k \geq 2$, $n \geq k$, $S \in \binom{[n]}{k}$ and $s \in S$ be given. Then, there exists a probability distribution $D = D^{(s,S,n)}$ over $\mathbf{S}_n$ such that:*

- *for each $\varnothing \subsetneq T \subseteq [n]$, $T \neq S$, $D_T = U_T$,*
- *$D_S(s) = \frac{1}{|S|} + \frac{1}{\beta_{n,k}}$,*

and hence $|D_S - U_S|_{\mathrm{tv}} \geq |D_S - U_S|_\infty \geq 1/\beta_{n,k}$.

Proof. By relabeling, without loss of generality, we assume $S = [k]$ and $s = 1$. Given a permutation π, we write $\underline{\pi}$ to denote the set of the k bottom elements of π. Let

$$c_i = 1 - (-1)^i \cdot \frac{\binom{k-1}{i-1}}{\binom{k-1}{\lfloor \frac{k-1}{2} \rfloor}}.$$

If $C = \sum_{i=1}^k c_i$, then note that

$$C = \sum_{i=1}^k c_i = k - \frac{1}{\binom{k-1}{i-1}} \cdot \sum_{i=1}^k \left((-1)^i \cdot \binom{k-1}{i-1}\right)$$
$$(3.2)= k,$$

using $\sum_{i=1}^k (-1)^i \cdot \binom{k-1}{i-1} = 0$.

We first define D explicitly:

$$D(\pi) = \begin{cases} 1/n! & \underline{\pi} \neq [k], \\ c_{\pi(1)}/n! & \text{otherwise.} \end{cases}$$

Observe that

$$P' = \sum_{\substack{\pi\in\mathbf{S}_n\\ \underline{\pi}\neq[k]}} D(\pi) = (n! - (n-k)!\cdot k!)\cdot\frac{1}{n!}$$

$$= 1 - \binom{n}{k}^{-1}, \tag{3.3}$$

and using (3.2),

$$P'' = \sum_{\substack{\pi\in\mathbf{S}_n\\ \underline{\pi}=[k]}} D(\pi) = \sum_{i=1}^{k}\frac{c_i}{n!}\cdot(k-1)!\cdot(n-k)!$$

$$= \frac{(k-1)!\cdot(n-k)!}{n!}\cdot\sum_{i=1}^{k} c_i = \binom{n}{k}^{-1}. \tag{3.4}$$

(3.3) + (3.4) shows that D is indeed a probability distribution.

Next, by symmetry, we have that if $1 \notin T$, then the projection of a random permutation $\pi \sim D$ on T is going to be uniform at random. It follows that for any $\varnothing \subset T \subseteq [n]$, $\Pr[\pi^*(T) = i] = \Pr[\pi^*(T) = j]$ for each $\{i,j\} \in \binom{T\setminus\{1\}}{2}$.

Thus, there only remain to be shown that: (i) $\Pr[\pi^*([k]) = 1] = \frac{1}{k} + \frac{1}{\beta_{n,k}}$ and (ii) $\Pr[\pi^*(T) = 1] = \frac{1}{|T|}$ for each T such that $T \neq [k]$ and $1 \in T$.

To prove (i), observe that

$$\Pr[\pi^*([k]) = 1] = \frac{1}{k}\cdot P' + \frac{c_k}{C}\cdot P''$$

$$= \frac{1}{k}\left(1 - \binom{n}{k}^{-1}\right) + \frac{1}{k}\left(1 + \binom{k-1}{\lfloor\frac{k-1}{2}\rfloor}^{-1}\right)\binom{n}{k}^{-1}$$

$$= \frac{1}{k} + \frac{1}{k}\cdot\frac{\lfloor\frac{k-1}{2}\rfloor!\cdot\lceil\frac{k-1}{2}\rceil!\cdot k!\cdot(n-k)!}{(k-1)!\cdot n!}$$

$$= \frac{1}{k} + \frac{1}{\beta_{n,k}}.$$

We move on to (ii). Suppose that $t = |T|$, $\ell = |T \cap ([k]\setminus\{1\})|$ and $u = t - 1 - \ell$. Now, if $u \geq 1$, we have that no permutation π such that $\underline{\pi} = [k]$ satisfies $\pi^*(T) = 1$, since there will always exists some element in T that π ranks higher than all the elements in $[k]$. Therefore, the sum making up $\Pr[\pi^*(T) = 1]$ is composed only of permutations π such that $\underline{\pi} \neq [k]$. But, each of those permutations has probability $1/n!$ and hence $\Pr[\pi^*(T) = 1] = 1/|T|$.

Last, we address the case $u = 0$. In that case, we have $\{1\} \subset T \subset [k]$, thus $0 \leq \ell \leq k-2$. Then,

$$\Pr[\pi^*(T) = 1] = \frac{1}{\ell+1}\cdot P' + \frac{\sum_{i=\ell+1}^{k}\left(c_i\cdot\binom{i-1}{\ell}\right)}{C\cdot\binom{k-1}{\ell}}\cdot P''.$$

We begin by solving the sum in the expression:

$$\sum_{i=\ell+1}^{k}\left(c_i\cdot\binom{i-1}{\ell}\right)$$

$$= \sum_{i=\ell+1}^{k}\left(\left(1 - (-1)^i\cdot\frac{\binom{k-1}{i-1}}{\binom{k-1}{\lfloor\frac{k-1}{2}\rfloor}}\right)\cdot\binom{i-1}{\ell}\right)$$

$$= \sum_{i=\ell+1}^{k}\binom{i-1}{\ell} + \binom{k-1}{\lfloor\frac{k-1}{2}\rfloor}^{-1}\cdot\sum_{i=\ell+1}^{k}\left((-1)^i\cdot\binom{k-1}{i-1}\cdot\binom{i-1}{\ell}\right)$$

$$= \binom{k}{\ell+1} + \binom{k-1}{\lfloor\frac{k-1}{2}\rfloor}^{-1}\cdot\sum_{i=\ell+1}^{k}\left((-1)^i\cdot\binom{k-1}{\ell}\cdot\binom{k-\ell-1}{i-\ell-1}\right)$$

$$= \binom{k}{\ell+1} + \binom{k-1}{\lfloor\frac{k-1}{2}\rfloor}^{-1}\cdot\binom{k-1}{\ell}\cdot\sum_{i=0}^{k-\ell-1}\left((-1)^{i+\ell+1}\cdot\binom{k-\ell-1}{i}\right)$$

$$= \binom{k}{\ell+1},$$

since $\sum_{i=0}^{k-\ell-1}\left((-1)^{i+\ell+1}\cdot\binom{k-\ell-1}{i}\right) = 0$. Using this, we finally get

$$\Pr[\pi^*(T) = 1] = \frac{1}{\ell+1}\cdot P' + \frac{\binom{k}{\ell+1}}{k\cdot\binom{k-1}{\ell}}\cdot P''$$

$$= \frac{1}{\ell+1}\cdot P' + \frac{\frac{k}{\ell+1}\cdot\binom{k-1}{\ell}}{k\cdot\binom{k-1}{\ell}}\cdot P'' = \frac{1}{\ell+1} = \frac{1}{|T|}. \quad \blacksquare$$

Observe that, as long as $k \leq \frac{n}{2}$, we have $\frac{1}{\beta_{n,k}} \geq \Omega\left(2^{-3n/2}\right)$. Moreover, for any fixed constant $k \geq 2$, we have $\frac{1}{\beta_{n,k}} = \Theta(n^{-k})$.

COROLLARY 3.1. *At least $\Omega(2^n)$ calls to* MAX-DIST *oracle are necessary in order to reconstruct, for all $\varnothing \subset T \subseteq [n]$, D_T to within a total variation error (or ℓ_∞-error) of $o\left(2^{-3n/2}\right)$.*

Proof. Suppose that D is chosen as follows: let S be a uniform at random subset of cardinality in the range $[\lfloor n/2\rfloor]\setminus\{1\}$ and let $s \in S$ be a uniform at random element of S. Let $D = D^{(s,S,n)}$ of Lemma 3.1. Then, an algorithm has to obtain the distributions of a constant fraction of the subsets of $[n]$ to get, with probability $\Omega(1)$, a maximum ℓ_∞-error (and hence total variation

error) of $o\left(2^{-3n/2}\right)$ for the distribution of each subset, in particular, for the distribution D_S of the unknown set S. ■

An analogous proof (omitted) can be used to show that, for each constant $k \geq 2$, $\Omega(n^k)$ queries are needed to get within a total variation (or ℓ_∞-) error of $o(n^{-k})$ from each true set distribution.

4 Algorithms for bounded support

Given the strong lower bound in Section 3, we turn to what is algorithmically possible with some reasonable assumptions. To this end, in this section we assume that the unknown distribution D has support on at most k permutations. For this important special case, we will give two algorithms that can be used together to solve the problem. Since the representation of a RUM as permutations need not be unique, so we do not seek to learn the permutations exactly. Instead, the first algorithm (Lemma 4.1) learns some properties of the permutations. Specifically, it inductively finds subsets that lie (in arbitrary order) at the head of at least one permutation, followed immediately by a specific element. There are only polynomially many such (set, element) pairs, and they can all be reconstructed in $O(nk)$ Max-Dist oracle accesses. The second algorithm (Theorem 4.1) then uses these probabilities to compute D_T for any T. In fact we will also show how one can use $(\epsilon^{-1}nk)^{O(1)}$ Max-Sample oracle queries to get an arbitrarily good approximations of D_T.

Lemma 4.1. *Suppose* $|\mathrm{supp}(D)| = k$. *Then, using* $O(nk)$ *calls to the* Max-Dist *oracle we can compute for each set* $S \subseteq [n]$, *and for each* $s \notin S$, *the probability* $P_{S,s}$ *that the returned permutation (i) has the elements of* S *(in an arbitrary order) in its first* $|S|$ *positions and (ii) has element* s *in its* $(|S|+1)$*st position.*

Proof. Observe that, for any $c = 0, \ldots, n-1$, there are at most k pairs (S, s) such that $|S| = c$ and $P_{S,s} > 0$. Indeed, if otherwise, $|\mathrm{supp}(D)| > k$.

We will prove the claim by induction on $c = |S|$; we will show that, given $P_{T,t}$ for each $|T| \leq c-1$, by using at most k Max-Dist oracle queries, we can compute $P_{S,s}$ for each $s \notin S$ such that $c = |S|$.

For the base case $c = 0$ (i.e., $S = \varnothing$), the claim is trivial. A a single Max-Dist oracle query on the full set will give us $P_{\varnothing,t}$ for each $t \in [n]$.

Now, suppose that the claim is true for $c - 1 \geq 0$. For each pair (T, t), such that $|T| = c-1$ and $P_{T,t} > 0$, we perform a Max-Dist oracle query on the set $[n] \setminus S$ for $S = T \cup \{t\}$. Let $M_{[n]\setminus S,x}$ be the probability that x wins in the set $[n] \setminus S$. Now, observe that:

- If there exists no pair (T, t) such that $S = T \cup \{t\}$ and such that $P_{T,t} > 0$, then $P_{S,s} = 0$.
- Instead, if there exists some pair (T, t) such that $S = T \cup \{t\}$, then we have:

$$P_{S,s} = M_{[n]\setminus S,s} - \sum_{T \subset S} P_{T,s}. \tag{4.5}$$

  Indeed, the probability that the elements of S are in the first $|S|$ positions and that s is in the $(|S|+1)$st position is equal to the probability that s is the first element in the subset $[n] \setminus S$, minus the probability that s ends up in some of the first $|S|$ positions.

Thus, we make $O(k)$ Max-Dist oracle queries for each $c = 0, \ldots, n-1$. The total number of queries is then $O(nk)$. ■

Corollary 4.1. *Using* $O(nk)$ *calls to an* $(\epsilon/nk, (nk)^{-2})$*-approx-*Max-Dist *oracle we can, with probability* $1 - o(1)$, *compute for each* $s \notin S$, *approximations* $\hat{P}_{S,s}$ *of* $P_{S,s}$ *such that* $|\hat{P}_{S,s} - P_{S,s}| = O(\epsilon)$.

Proof. Define $\hat{P}_{S,s} = \hat{M}_{[n]\setminus S,s} - \sum_{T \subset S} \hat{P}_{T,s}$, where $\hat{M}_{[n]\setminus S,s}$ is obtained from an $(\epsilon/nk, (nk)^{-2})$-approx-Max-Dist oracle. Now, if we unwind the recursive expression of $\hat{P}_{S,s}$ in (4.5), stating it only in terms of the $\hat{M}_{S,s}$'s, we obtain the following. (i) Each $\hat{M}_{S,s}$ will have a coefficient bounded between $[-1, -1]$; (ii) $M_{S,s} = 0$ implies $\hat{M}_{S,s} = 0$; (iii) the number of non-zero $M_{S,s}$ can be upper bounded by $O(nk)$, and (iv) $|\hat{M}_{S,s} - M_{S,s}| \leq \epsilon/nk$. Thus, we have $|\hat{P}_{S,s} - P_{S,s}| \leq O(\epsilon)$. ■

Lemma 4.1 and Corollary 4.1 immediately yield the following.

Theorem 4.1. *Suppose* $|\mathrm{supp}(D)| = k$. *Using* $O(nk)$ Max-Dist *oracle queries, we can compute* D_T *for any set* $T \subseteq [n]$. *Using* $O(\epsilon^{-2}n^5k^2\log(kn))$ Max-Sample *oracle queries, we can compute with probability* $1 - o(1)$, *for any* $T \subseteq [n]$, *a distribution* $\tilde{D}_T$ *such that* $|\tilde{D}_T - D_T|_{\mathrm{tv}} \leq O(\epsilon)$.

Proof. Using $O(nk)$ Max-Dist oracle queries, we compute $P_{S,s}$, for all $S \subseteq [n]$ and for all $t \in [n] \setminus S$. Now, let $\varnothing \neq T \subseteq [n]$ be a subset and let $t \in T$. We have:

$$\Pr[\pi^*(T) = t] = \sum_{S \subseteq [n]\setminus T\setminus\{t\}} P_{S,t}.$$

As in the proof of Corollary 4.1, we note that $\sum_{S\subseteq[n]\setminus T\setminus\{t\}} \hat{P}_{S,t}$ is composed of at most $O(nk)$ non-zero $\hat{M}_{S,s}$'s. Thus, an $(\epsilon/(n^2k), (nk)^{-2})$-approx-Max-Dist oracle would give us an approximation error of ϵ/n for $\Pr[\pi^*(T) = t]$, since $|T| \leq n$. Claim 2.1 guarantees that an $(\epsilon/(n^2k), (nk)^{-2})$-approx-Max-Dist oracle can be realized with $O(\epsilon^{-2}n^5k^2\log(nk))$ queries to a Max-Sample oracle. ■

4.1 On learning the rank of an element Having established our algorithm for RUMs with support on at most k permutations, one may ask whether the proof techniques employed here generalize to larger support. We offer some negative evidence regarding such extensions.

Recall that $P_{S,s}$, for $S \subset [n]$ and $s \in [n] \setminus S$, is the probability that a random permutation has the elements of S in its first $|S|$ positions in any order, and that it has element s in its $(|S|+1)$st position. In the proof of Lemma 4.1 we have demonstrated an observation that we now make explicit.

OBSERVATION 4.1. $P_{S,s} = D_{[n]\setminus S}(s) - \sum_{T \subset S} P_{T,s}$.

With this observation, we could learn, given sufficiently many queries, the distribution of the rank of an element in the random permutation. Indeed, the probability that element s ends up in position $k+1$ is then equal to $\sum_{S \in \binom{[n]\setminus\{s\}}{k}} P_{S,s}$.

In contrast, we now show that learning the distribution of the rank of a specific element s in the random permutation requires $\Omega(2^n)$ MAX-DIST oracle queries.

THEOREM 4.2. *An (adaptive) algorithm needs $\Omega(2^n)$ MAX-DIST oracle queries to distinguish whether an element appears with probability 1 in odd-ranked positions or appears with probability 1 in even-ranked positions.*

Proof. For simplicity, let $s = 1$. Consider the following two processes to generate different distributions over permutations:

- In process $\mathcal{E}$, (i) select a subset $S \subseteq \{2, \ldots, n\}$ u.a.r of *even* cardinality; (ii) permute S u.a.r. obtaining a permutation π_h; (iii) permute $\{2, \ldots, n\} \setminus S$ u.a.r. obtaining permutation π_t; (iv) return the permutation $\pi_h \cdot (1) \cdot \pi_t$.
- In process $\mathcal{O}$, (i) select a subset $S \subseteq \{2, \ldots, n\}$ of *odd* cardinality; and follow (ii)–(iv) as in process $\mathcal{E}$.

By symmetry, in both processes $\mathcal{E}$ and $\mathcal{O}$, the max distribution of any set S will give the same probability to each element of $S \setminus \{s\}$. Moreover, if $s \in S$ and S contains exactly c other elements (so that $|S| = c+1$), then (i) if $c < n-1$, then both $\mathcal{E}$ and $\mathcal{O}$ will give probability 2^{-c} to s and (ii) if $c = n-1$ (so that $S = [n]$), then $\mathcal{E}$ will choose element s with probability 0, while $\mathcal{O}$ will choose element s with probability 2^{2-n}.

The algorithm, to distinguish between $\mathcal{E}$ and $\mathcal{O}$, then has to perform $\Omega(2^n)$ MAX-DIST oracle queries on the slate $[n]$. Moreover, since the supports of the distributions of the rank of s in $\mathcal{E}$ and $\mathcal{O}$ are disjoint, the algorithm requires $\Omega(2^n)$ MAX-DIST oracle queries to learn anything about the distribution of s to within a total variation error smaller than 1. ■

5 Queries with bounded subset sizes

We now turn to the situation in which oracle queries are possible only for subsets of bounded size. This situation arises naturally when it is not reasonable to offer the user an enormous slate of options. We first show in Section 5.1 that even super-polylogarithmic bounds on subset size are limiting enough to force any algorithm to make worst-case statistical error on some sets. On the flip side, we then show in Section 5.2 that if the queried subsets are of a certain size, then one can reasonably approximate D_S for subsets S of slightly larger size.

5.1 High-distance lower bounds We first show that if the queried subsets have size at most $\lg n$ (improved later to $2^{\Theta(\sqrt{\log n})}$), one cannot avoid making the maximum possible statistical error while reconstructing D_T for some T. We use an intriguing connection: these lower bounds are based on the lower bounds of [4,9] for the k-deck reconstruction problem.

We begin by stating our first lower bound.

THEOREM 5.1. *Let $n = 2^i$ for $i \in \mathbb{Z}^+$. Then, there are two distributions D, D' over $\mathbf{S}_n$ such that:*

- *for each set $Q \subseteq [n]$, with $|Q| \leq \lg n$, it holds $D_Q = D'_Q$ and*
- *$\operatorname{supp}\left(D_{[n]}\right) \cap \operatorname{supp}\left(D'_{[n]}\right) = \varnothing$ and thus $\left|D_{[n]} - D'_{[n]}\right|_{\mathrm{tv}} = 1$.*

The distributions D and D' that we are going to use are based on the *Prouet–Thue–Morse* (PTM) sequence $\{s_i\}_{i=0}^{\infty}$. Let $s_0 = (1)$, and $s_{i+1} = s_i \cdot (-s_i)$, where $\cdot$ is the concatenation operator, and $-x$ is the complement of x (i.e., if $x = (x_1, \ldots, x_n)$ then $-x = (-x_1, \ldots, -x_n)$). For example, the first four terms of the PTM sequence are: $s_0 = (1)$, $s_1 = (1, -1)$, $s_2 = (1, -1, -1, 1)$ and $s_3 = (1, -1, -1, 1, -1, 1, 1, -1)$.

Given a sequence (or a string) s, the *k-deck* of s is the multiset of k-subsequences of s. Thus, the k-deck of s has cardinality $\binom{|s|}{k}$. We use the following result.

THEOREM 5.2. (MANVEL ET AL. [9]) *For each $i \geq 0$, and for each $0 \leq k \leq i$, the k-deck of s_i equals the k-deck of $-s_i$.*

We can now prove Theorem 5.1:

Proof. We will use the following two distributions D and D'. For D, $\operatorname{supp}(D)$ will be equal to the set of permutations of $[n]$ that have the elements of $[n/2]$ in the positions where s_i has value 1. Likewise, for D', $\operatorname{supp}(D')$ will be equal to the set of permutations of $[n]$ that have the elements of $[n/2]$ in the positions where s_i has value -1. Both D and D' will choose u.a.r. in their supports. (For example, if $i = 2$, D will choose u.a.r.

in $\{(1,3,4,2),(2,3,4,1),(1,4,3,2),(2,4,3,1)\}$ and D' will choose u.a.r. in $\{(3,1,2,4),\ (3,2,1,4),\ (4,1,2,3),\ (4,2,1,3)\}$.)

First, observe that by construction, $\operatorname{supp}\left(D_{[n]}\right) = [n/2]$ and $\operatorname{supp}\left(D'_{[n]}\right) = [n] \setminus [n/2]$, and hence the supports of $D_{[n]}$ and $D'_{[n]}$ are disjoint.

Next, consider any set $Q \subseteq [n]$ such that $|Q| \leq \lg n$. We give a bijection f between $\operatorname{supp}(D)$ and $\operatorname{supp}(D')$ that guarantees that the relative ordering of the elements of Q remains equal in π and $f(\pi)$. This directly implies that the probability that the distribution of the maximum element of Q does not change from D to D'.

By Theorem 5.2, there exists a bijection b between the $|Q|$-tuples of indices of s_i and the $|Q|$-tuples of indices of $-s_i$ that preserves ordering and values.

Now, consider some $\pi \in \operatorname{supp}(D)$. Let x be the tuple containing the $|Q|$ indices of the elements of Q in π. The permutation $f(\pi)$ will have the elements of Q, in the same order of π, in the indices $b(x)$. Assign the elements of $[n/2] \setminus Q$, in sorted order, to the unoccupied positions of $f(\pi)$ whose corresponding value in s_i is 1, and assign the elements of $([n] \setminus [n/2]) \setminus Q$, in sorted order, to the unoccupied positions of $f(\pi)$ whose corresponding value in s_i is -1. Observe: (i) $f(\pi) \in \operatorname{supp}(D')$; (ii) $f^{-1}(f(\pi)) = \pi$; and (iii) π and $f(\pi)$, when restricted to Q, are equal.

Since $|\operatorname{supp}(D)| = |\operatorname{supp}(D')|$ and since D and D' choose uniformly at random in their respective supports, we have that $D_Q = D'_Q$. ■

Finally, we show the following stronger lower bound, built in a manner equivalent to the one above, but using the stronger k-deck reconstruction lower bound of Dudík and Schulman [4]; we omit the proof in this version.

THEOREM 5.3. *For each n in an increasing infinite sequence, there exists two probability distributions D and D' over permutations of $[n]$, such that:*

- *for each set $Q \subseteq [n]$, with $|Q| \leq 2^{\Theta\left(\sqrt{\log n}\right)}$, it holds $D_Q = D'_Q$;*
- $\operatorname{supp}\left(D_{[n]}\right) \cap \operatorname{supp}\left(D'_{[n]}\right) = \varnothing$, $\left|D_{[n]} - D'_{[n]}\right|_{\text{tv}} = 1$.

5.2 Approximating large subsets with small subset queries We now show that, if one has access to D_T for T's of a certain size, one can get reasonable approximations to D_S for S's that are not much larger.

THEOREM 5.4. *For any $K > k$, if we are given D_T for all $|T| = k$ then for any set $S, |S| = K$, we can find $\tilde{D}_S$ such that $|\tilde{D}_S - D_S|_{\text{tv}} \leq 1 - \frac{k}{K}$.*

Proof. Let $D_T(i)$ be the probability of i in D_T. The algorithm we will employ is simple: for a given set S, with $|S| = K$, return the distribution

$$A_S^{(k)} = \binom{K}{k}^{-1} \sum_{T \in \binom{S}{k}} D_T.$$

We have,

$$\begin{aligned}
&D_S(i) - A_S^{(k)}(i) \\
&= \sum_{\pi} \left(D(\pi) \cdot \left([\pi^*(S) = i] - \frac{\sum_{T \in \binom{S}{k}} [\pi^*(T) = i]}{\binom{K}{k}} \right) \right) \\
&= \sum_{\pi | \pi^*(S) = i} \left(D(\pi) \cdot \left(1 - \frac{\sum_{T \in \binom{S}{k}} [\pi^*(T) = i]}{\binom{K}{k}} \right) \right) \\
&\quad - \sum_{\pi | \pi^*(S) \neq i} \left(D(\pi) \cdot \frac{\sum_{T \in \binom{S}{k}} [\pi^*(T) = i]}{\binom{K}{k}} \right) \\
&= \sum_{\pi | \pi^*(S) = i} \left(D(\pi) \cdot \left(1 - \frac{\binom{K-1}{k-1}}{\binom{K}{k}} \right) \right) \\
&\quad - \sum_{\pi | \pi^*(S) \neq i} \left(D(\pi) \cdot \frac{\sum_{T \in \binom{S}{k}} [\pi^*(T) = i]}{\binom{K}{k}} \right) \\
&= \left(1 - \frac{\binom{K-1}{k-1}}{\binom{K}{k}} \right) \cdot \Pr\left[\pi^*(S) = i\right] \\
&\quad - \sum_{\pi | \pi^*(S) \neq i} \left(D(\pi) \cdot \frac{\sum_{T \in \binom{S}{k}} [\pi^*(T) = i]}{\binom{K}{k}} \right) \\
&= \left(1 - \frac{k}{K} \right) \cdot \Pr\left[\pi^*(S) = i\right] \\
&\quad - \sum_{\pi | \pi^*(S) \neq i} \left(D(\pi) \cdot \frac{\sum_{T \in \binom{S}{k}} [\pi^*(T) = i]}{\binom{K}{k}} \right) \\
&\leq \left(1 - \frac{k}{K} \right) \cdot \Pr\left[\pi^*(S) = i\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{i \in S} \max\left(0, D_S(i) - A_S^{(k)}(i)\right) \\
&\quad \leq \sum_{i \in S} \left(\left(1 - \frac{k}{K} \right) \cdot \Pr\left[\pi^*(S) = i\right] \right) \ = \ 1 - \frac{k}{K}.
\end{aligned}$$

Since both D_S and $A_S^{(k)}$ are probability distributions, we have

$$\begin{aligned}
&\sum_{i} \max\left(0, D_S(i) - A_S^{(k)}(i)\right) \\
&\quad = \sum_{i} \max\left(0, A_S^{(k)}(i) - D_S(i)\right) \ = \ |A_S^{(k)} - D_S|_{\text{tv}}.
\end{aligned}$$

The proof for an (ϵ, n^{-2K})-approx-MAX-DIST oracle is analogous and is omitted. ■

Thus, we have the following.

COROLLARY 5.1. *For any constant c and for any K, by using D_T for $|T| \leq K - c$, we can find D'_S for $|S| \leq K$ such that $|D'_S - D_S|_{\text{tv}} \leq c/K$. In particular, for any $0 \leq \epsilon \leq 1$, with less than $n^{(1-\epsilon)K}$ MAX-DIST oracle queries (to all sets of size at most $(1-\epsilon)K$), we can get approximate distributions for all sets of size at most K, each within total variation distance ϵ.*

It is easy to show (proof omitted) that, if one queries an (ϵ, n^{-2K})-approx-MAX-DIST oracle on all subsets of size k then, for each set subset S of size K, one can return a distribution that is at total variation distance $1-\frac{k}{K}+O(\epsilon)$ from D_S with probability at least $1-n^{-2K}$. The algorithm for this task is the same one as in the proof of Theorem 5.4. We recall that one can query a (ϵ, n^{-2K})-approx-MAX-DIST oracle on each set of size k with $O\left(\epsilon^{-2} \cdot k \cdot K \cdot \log n\right)$ calls to a MAX-SAMPLE oracle.

Finally, we show that our analysis is tight.

LEMMA 5.1. *Let D be the distribution that places all its mass on the identity permutation of $[K]$. Then, for each $k < K$, it holds that $\left|D_{[K]} - A^{(k)}_{[K]}\right|_{\text{tv}} = 1 - \frac{k}{K}$.*

Proof. Observe that $D_{[K]}(1) = 1$, and $D_{[K]}(i) = 0$ for each $i \geq 2$. On the other hand, we have that $A^{(k)}_{[K]}(i) = 0$ for each $i \geq K - k + 2$, and

$$A^{(k)}_{[K]}(i) = \frac{\binom{K-i}{k-1}}{\binom{K}{k}} = k \cdot \frac{(K-i)!(K-k)!}{K!(K-i-k+1)!},$$

for each $i \leq K - k + 1$, in particular, $A^{(k)}_{[K]}(1) = k/K$. Therefore, we have that

$$\begin{aligned}
&\left|D_{[K]} - A^{(k)}_{[K]}\right|_1 \\
&= \left(1 - \frac{k}{K}\right) + k \sum_{i=2}^{K-k+1} \frac{(K-i)!(K-k)!}{K!(K-i-k+1)!} \\
&= 2\left(1 - \frac{k}{K}\right). \quad \blacksquare
\end{aligned}$$

6 Representability

In this section we consider alternate and compact representations for distributions on permutations, with respect to our setting. We start by proving a coreset-type result: any distribution D on permutations can be transformed into a small-support distribution D' such that for all subsets T, D_T and D'_T are close in total variation distance. We assume that we have access to D.

THEOREM 6.1. *Let $0 < \epsilon < 1$. There exists a polynomial time algorithm that, given any distribution D on $\mathbf{S}_n$, is able to construct a distribution D' on $\mathbf{S}_n$ with $\text{supp}(D') = O(n^2\epsilon^{-2})$ such that, with probability $1-o(1)$, for each $\varnothing \neq S \subseteq [n]$ it holds $|D_S - D'_S|_{\text{tv}} \leq \epsilon$.*

Proof. Since we have access to D, we can take $T = 3n^2\epsilon^{-2}$ independent samples (i.e., permutations) $\pi_1, \ldots, \pi_T$ from D, and we let D' be the uniform distribution on the multiset of these samples, i.e., D' will choose $i \in [T]$ uniformly at random, and will return π_i.

Consider any set $\varnothing \neq S \subseteq [n]$, and let $s \in S$. Observe that $D'_S(s)$, which is the probability that $s \in S$ is the maximum element in the subset S with distribution D', is then a random variable. Clearly, $E\left[D'_S(s)\right] = D_S(s)$. Suppose that $D_S(s) \geq \frac{1}{n}$. Then,

$$\begin{aligned}
&\Pr\left[|D'_S(s) - D_S(s)| \geq \epsilon D_S(s)\right] \leq 2e^{-\frac{\epsilon^2}{3}\left(T \cdot E\left[D'_S(s)\right]\right)} \\
&\quad = \quad 2e^{-\frac{\epsilon^2}{3}\left(3n^2\epsilon^{-2}D_S(s)\right)} \quad \leq \quad 2e^{-n}
\end{aligned}$$

Now, if $D_S(s) = 0$, then necessarily $D'_S(s) = 0$. We can then assume $0 < D_S(s) \leq \frac{1}{n}$. Then,

$$\begin{aligned}
&\Pr\left[|D'_S(s) - D_S(s)| \geq \frac{\epsilon}{n}\right] \\
&\quad = \quad \Pr\left[|D'_S(s) - D_S(s)| \geq \frac{\epsilon}{n \cdot D_S(s)} \cdot D_S(s)\right] \triangleq p^\star.
\end{aligned}$$

Now, if $\frac{\epsilon}{n} \leq D_S(s) \leq \frac{1}{n}$, we can apply a standard Chernoff bound (if $X_1, \ldots, X_m$ are iid binary variables, and $X = \sum_{i=1}^m X_i$, then $\Pr\left[|X - E[X]| \geq \delta E[X]\right] \leq 2e^{-\frac{\delta^2}{3}E[X]}$, for each $0 < \delta \leq 1$), to bound

$$\begin{aligned}
p^\star \quad &\leq \quad 2e^{-\frac{\epsilon^2}{3n^2 D_S(s)^2} \cdot (T \cdot D_S(s))} \\
&= \quad 2e^{-\frac{\epsilon^2}{3n^2 D_S(s)} \cdot T} \quad = \quad 2e^{-\frac{1}{D_S(s)}} \quad \leq \quad 2e^{-n}.
\end{aligned}$$

If, instead, $D_S(s) \leq \frac{\epsilon}{n}$, we can apply a large-gap Chernoff bound (if $X_1, \ldots, X_m$ are iid binary variables, and $X = \sum_{i=1}^m X_i$, then $\Pr\left[|X - E[X]| \geq \delta E[X]\right] \leq 2e^{-\frac{\delta}{3}E[X]}$, for each $\delta \geq 1$), to bound

$$p^\star \leq 2e^{-\frac{\epsilon}{3nD_S(s)} \cdot (T \cdot D_S(s))} = 2e^{-\frac{\epsilon}{3n} \cdot T} = 2e^{-n/\epsilon} \leq 2e^{-n}.$$

Thus, for any given $\varnothing \neq S \subseteq [n]$, and any $s \in S$, we have that

$$\Pr\left[|D'_S(s) - D_S(s)| \geq \epsilon \cdot \max\left(\frac{1}{n}, D_S(s)\right)\right] \leq 2e^{-n}.$$

Applying a union bound over all $\varnothing \neq S \subseteq [n]$, and $s \in S$, we get that

$$\begin{aligned}
&\Pr\left[\exists s, S : |D'_S(s) - D_S(s)| \geq \epsilon \cdot \max\left(\frac{1}{n}, D_S(s)\right)\right] \\
&\quad \leq \quad 2e^{-n} \cdot n2^n = o(1).
\end{aligned}$$

Thus, with probability $1 - o(1)$, we have that for all $\varnothing \subset S \subseteq [n]$,

$$|D_S - D'_S|_1 \leq \frac{\epsilon}{n}|S| + \epsilon \sum_{s \in S} D_S(s) \leq \frac{\epsilon}{n} \cdot n + \epsilon \cdot 1 = 2\epsilon. \quad \blacksquare$$

We next show that distribution on permutations can be approximated by distribution over depth-1 trees. Recall the choice process for depth-1 trees. Let T be a distribution over depth-1 trees $t_1, t_2, ...$, where the tree t_i is a star on $[n]$ leaves and the weight of the edge from the root to the jth leaf is w_{ij}. The choice process first chooses a tree t_i according to T, and then chooses an element j with probability $w_{ij} / \sum_{j' \in [n]} w_{ij'}$.

THEOREM 6.2. *For any distribution D on permutations, there exists a distribution T over depth-1 trees such that for each $\varnothing \neq S \subseteq [n]$, we have $|D_S - T_S|_{\text{tv}} \leq \epsilon$.*

Proof. For each permutation $\pi \in \mathbf{S}_n$ we create a depth-1 tree t_π and give it weight $T(t_\pi) = D(\pi)$ in the trees. The tree t_π will be a star with n leaves, one for each element in $[n]$. For each $i \in [n]$, the weight of the edge that connects the root of t_π to the leaf corresponding to the element i is $\epsilon^{-\pi(i)}$.

Now, for any $\varnothing \subset S \subseteq [n]$, $t_\pi(S)$ will be a distribution giving weight $\epsilon^{-\pi(i)}$ to element $i \in S$. It follows that the element of S having maximum index in π will have, in $t_\pi(S)$, a probability that is larger, by a factor of at least $\Omega\left(\epsilon^{-1}\right)$, than the sum of probabilities of the other elements in S. Thus, the probability that $t_\pi(S)$ will return the element of largest index in π is at least $1 - \epsilon$. The proof is complete. $\blacksquare$

6.1 Impossibility of learning the distribution over permutations Note that the above two compact representation reductions require an explicit access to the distribution D on permutations. One might wonder if a powerful oracle such as the MAX-DIST oracle might allow us to construct the distribution D. However, the answer turns out to be negative: we show that it is impossible to construct the distribution over permutations based only on the MAX-DIST oracle.

LEMMA 6.1. *For each $n \geq 4$, there exists two distribution D, D' over $\mathbf{S}_n$ such that:*

- *for each $\varnothing \subsetneq S \subseteq [n]$, it holds $D_S = D'_S$*
- *$|D - D'|_{\text{tv}} = 1$.*

Moreover, $|\text{supp}(D)| = |\text{supp}(D')| = 2$.

Proof. Let $n \geq 4$ be given; we will create two distributions D and D' with disjoint supports and such that each permutation in the support of D or of D' will contain the elements of $\{1, 2, 3, 4\}$ in some order in its topmost 4 positions, and the remaining elements naturally ordered in the remaining positions. The distributions are defined as follows:

- D assigns probability $1/2$ to the permutation $(\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, 5, 6, \ldots, n-1, n)$ and probability $1/2$ to the permutation $(\mathbf{2}, \mathbf{1}, \mathbf{4}, \mathbf{3}, 5, 6, \ldots, n-1, n)$;
- D' assigns probability $1/2$ to the permutation $(\mathbf{2}, \mathbf{1}, \mathbf{3}, \mathbf{4}, 5, 6, \ldots, n-1, n)$ and probability $1/2$ to the permutation $(\mathbf{1}, \mathbf{2}, \mathbf{4}, \mathbf{3}, 5, 6, \ldots, n-1, n)$.

Since the supports of D and D' are disjoint, we have $|D - D'|_{\text{tv}} = 1$. We will show below that, for each set $\varnothing \subsetneq S \subseteq [n]$, it holds $D_S = D'_S$. This will imply that it is impossible to reconstruct the unknown distribution over permutations by only using the MAX-DIST oracle.

Indeed, let $\varnothing \subseteq S \subseteq [n]$ be any set. If $\{1, 2\} \subseteq S$, then both D_S and D'_S are uniform on $\{1, 2\}$; if, instead, $|\{1, 2\} \cap S| = 1$ then both $D(S)$ and $D'(S)$ give probability 1 to the one element in $\{1, 2\} \cap S$. We can then assume that $\{1, 2\} \cap S = \varnothing$; we proceed to consider sets S that intersect $\{3, 4\}$. Again, if $\{3, 4\} \subseteq S$, then both D_S and D'_S are uniform on $\{3, 4\}$; if $|\{3, 4\} \cap S| = 1$, then both D_S and D'_S give probability 1 to the one element in $\{3, 4\} \cap S$. Finally, we can assume that $\{1, 2, 3, 4\} \cap S = \varnothing$. In this case, both D_S and D'_S give probability 1 to $\min_{i \in S} i$. $\blacksquare$

7 Conclusions

In this paper we studied a version of the permutation reconstruction problem, motivated by algorithmic questions in discrete choice theory. We believe that this topic has of untapped research potential and immense practical value. Our work, while providing a conceptual starting point, barely scratches the surface in terms of what could be studied computationally. The power and limitations of the MAX-SAMPLE oracle and the MAX-DIST oracle still need to be fully understood. For instance, the following obvious question stands out: how much can adaptivity help in approximate reconstruction?

References

[1] A. Ammar, S. Oh, D. Shah, and L. F. Voloch. What's your choice?: Learning the mixed multi-nomial. *SIGMETRICS Perform. Eval. Rev.*, 42(1):565–566, 2014.

[2] A. R. Benson, R. Kumar, and A. Tomkins. On the relevance of irrelevant alternatives. In *WWW*, pages 963–973, 2016.

[3] J. Blanchet, G. Gallego, and V. Goyal. A Markov chain approximation to choice modeling. In *EC*, pages 103–104, 2013.

[4] M. Dudík and L. J. Schulman. Reconstruction from subsequences. *Journal of Combinatorial Theory, Series A*, 103(2):337–348, 2003.

[5] V. F. Farias, S. Jagabathula, and D. Shah. A data-

driven approach to modeling choice. In *NIPS*, pages 504–512, 2009.

[6] A. Gupta and D. Hsu. Parameter identification in Markov chain choice models. Technical Report 1706.00729, arXiv, 2017.

[7] N. Kallus and M. Udell. Revealed preference at scale: Learning personalized preferences from assortment choices. In *EC*, pages 821–837, 2016.

[8] R. D. Luce. On the possible psychophysical laws. *Psychological Review*, 66(2):81, 1959.

[9] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, and P. Stockmeyer. Reconstruction of sequences. *Discrete Mathematics*, 94(3):209–219, 1991.

[10] J. Marschak. Binary choice constraints on random utility indications. In K. Arrow, editor, *Stanford Symposium on Mathematical Methods in the Social Sciences*, pages 312–329. Stanford University Press, Stanford, CA, 1960.

[11] D. McFadden and K. Train. Mixed MNL models of discrete response. *Journal of Applied Econometrics*, 15:447–470, 2000.

[12] S. Oh and D. Shah. Learning mixed multinomial logit model from ordinal data. In *NIPS*, pages 595–603, 2014.

[13] K. E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.