2018 IEEE 59th Annual Symposium on Foundations of Computer Science

Efficiently Learning Mixtures of Mallows Models

Allen X. Liu
Math
Massachusetts Institute of Technology
Cambridge, USA
cliu568@mit.edu

Ankur Moitra
Math, CSAIL and IDSS
Massachusetts Institute of Technology
Cambridge, USA
moitra@mit.edu

***Abstract*—Mixtures of Mallows models are a popular generative model for ranking data coming from a heterogeneous population. They have a variety of applications including social choice, recommendation systems and natural language processing. Here we give the first polynomial time algorithm for provably learning the parameters of a mixture of Mallows models with any constant number of components. Prior to our work, only the two component case had been settled. Our analysis revolves around a determinantal identity of Zagier [1] which was proven in the context of mathematical physics, which we use to show polynomial identifiability and ultimately to construct test functions to peel off one component at a time.**

To complement our upper bounds, we show information-theoretic lower bounds on the sample complexity as well as lower bounds against restricted families of algorithms that make only local queries. Together, these results demonstrate various impediments to improving the dependence on the number of components. They also motivate the study of learning mixtures of Mallows models from the perspective of beyond worst-case analysis. In this direction, we show that when the scaling parameters of the Mallows models have separation, there are much faster learning algorithms.

I. Introduction

A. Background

User preferences — in a wide variety of settings ranging from voting [2] to information retrieval [3] — are often modeled as a distribution on permutations. Here we study the problem of learning a mixture of Mallows models from random samples. First, a Mallows model $M(\phi, \pi^*)$ is described by a center π^* and a scaling parameter ϕ. The probability of generating a permutation π is

$$\Pr_{M(\phi,\pi^*)}[\pi] = \frac{\phi^{d_{KT}(\pi,\pi^*)}}{Z}$$

where d_{KT} is the Kendall-Tau distance [4] and Z is a normalizing constant that only depends on ϕ and the number of elements being permuted which we denote by n. C. L. Mallows introduced this model in 1957 and gave an inefficient procedure for sampling from them: Rank every pair of elements randomly and independently so that they agree with π^* with probability $\frac{1}{1+\phi}$ and output the ranking if it is a total ordering. Doignon et al. [5] discovered a repeated insertion-based model that they proved is equivalent to the Mallows model and more directly lends itself to an efficient sampling procedure.

The Mallows model is a natural way to represent noisy data when there is one true ranking that correlates well with everyone's own individual ranking. However in many settings (e.g. voting [6], recommendation systems) the population is heterogenous and composed of two or more subpopulations. In this case, it is more appropriate to model the data as a mixture of simpler models. Along these lines, there has been considerable interest in fitting the parameters of a mixture of Mallows models to ranking data [7], [8], [9]. However most of the existing approaches (e.g. Expectation-Maximization [7]) are heuristic and only recently were the first algorithms with provable guarantees given. For a single Mallows model, Braverman and Mossel [10] showed how to learn it by quantifying how close the empirical average of the ordering of elements is to the ordering given by π^* as the number of samples increases.

Awasthi et al. [11] gave the first polynomial time algorithm for learning mixtures of two Mallows models. Their algorithm learns the centers π_1 and π_2 exactly and the mixing weights and scaling parameters up to an additive θ with running time and sample complexity

$$\text{poly}(n, \frac{1}{w_{min}}, \frac{1}{\phi_1(1-\phi_1)}, \frac{1}{\phi_2(1-\phi_2)}, \frac{1}{\theta})$$

Here ϕ_1 and ϕ_2 are the scaling parameters and w_{min} is the smallest mixing weight. Their algorithm works based on recasting the parameter learning problem in the language of tensor decompositions, similarly to other algorithms for learning latent variable models [12]. However there is a serious complication in that most of the entries in the tensor are exponentially small. So even though we can compute unbiased estimates of

2575-8454/18/$31.00 ©2018 IEEE
DOI 10.1109/FOCS.2018.00066

the entries of a tensor whose low rank decomposition would reveal the parameters of the Mallows model, most of the entries cannot be meaningfully estimated from few samples. Instead, Awasthi et al. [11] show how the entries that can be accurately estimated can be used to learn the prefixes of the permutations, which can be bootstrapped to learn the rest of the parameters. In fact before their work, it was not even known whether a mixture of two Mallows models was *identifiable* — i.e. whether its parameters can be uniquely determined from an infinite number of samples.

The natural open question was to give provable algorithms for learning mixtures of any constant number of Mallows models. For other learning problems like mixtures of product distributions [13], [14] and mixtures of Gaussians [15], [16], [17], algorithms for learning mixtures of two components under minimal conditions were eventually extended to any constant number of components. Chierichetti et al. [18] showed that when the number of components is exponential in n, identifiability fails. On the other hand, when all the scaling parameters are the same and known, Chierichetti et al. [18] showed that it is possible to learn the parameters when given an arbitrary (and at least exponential in $\binom{n}{2}$) number of samples. Their approach was based on the Hadamard matrix exponential. They also gave a clustering-based algorithm that runs in polynomial time and works whenever the centers are well-separated according to the Kendall-Tau distance, by utilizing recent concentration bounds for Mallows models that quantify how close a sample π is likely to be to the center π^* [19].

B. Our Results and Techniques

Our main result is a polynomial time algorithm for learning mixtures of Mallows models for any constant number of components. Let d_{TV} be the total variation distance, let w_{min} be the smallest mixing weight, and let U denote the uniform distribution over the $n!$ possible permutations. We prove:

Theorem I.1. *For any constant k, given samples from a mixture of k Mallows models*

$$M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$$

where $d_{TV}(M(\phi_i, \pi_i), M(\phi_j, \pi_j)) \geq \mu$ for all $i \neq j$, $d_{TV}(M(\phi_i, \pi_i), U) \geq \mu$ for all i and $n \geq 10k^2$, there is an algorithm whose running time and sample complexity are

$$\textit{poly}(n, \frac{1}{w_{min}}, \frac{1}{\mu}, \frac{1}{\theta}, \log \frac{1}{\delta})$$

for learning each center π_i exactly and the mixing weights and scaling parameters to within an additive θ. Moreover the algorithm succeeds with probability at least $1 - \delta$.

A main challenge in learning mixtures of Mallows models is in establishing *polynomial identifiability* — i.e. that the parameters of the model can be approximately determined from a polynomial number of samples. When addressing this question, there is a natural special matrix A to consider: Let A be an $n! \times n!$ matrix whose rows and columns are indexed by permutations with

$$A_{\pi,\sigma} = \phi^{d_{KT}(\pi,\sigma)}$$

Zagier [1] used tools from representation theory to find a simple expression for the determinant of A. Interestingly his motivation for studying this problem came from interpolating between Bose and Fermi statistics in mathematical physics. We can translate his result into our context by observing that the columns of A, after normalizing so that they sum to one, correspond to Mallows models with the same fixed scaling parameter ϕ. Thus Zagier's result implies that any two distinct mixtures M and M' of Mallows models, where all the components have the same scaling parameter, produce different distributions[1].

However the quantitative lower bounds that follow from Zagier's expression for the determinant are too weak for our purposes, and are not adapted to the number of components in the mixture. We exploit symmetry properties to show lower bounds on the length of any column of A projected onto the orthogonal complement of any other $k-1$ columns, which allows us to show that not only does A have full rank, any small number of its columns are robustly linearly independent [20]. More precisely we prove:

Theorem I.2. *Let $\phi < 1 - \epsilon$. Let $c_1, c_2, \cdots, c_k$ be any k distinct columns of A, normalized so that they each sum to one. Then*

$$\|z_1 c_1 + \cdots + z_k c_k\|_1 \geq \frac{\max_i(|z_i|)\epsilon^{2k^2}}{2n^{4k}(k+1)^{2k^2+4k}}$$

where z_i are arbitrary real coefficients.

Even though this result nominally applies to mixtures of Mallows models where all the scaling parameters are the same, we are able to use it as a black box to solve the more general learning problem. We reformulate

[1]This result was rediscovered by Chierichetti et al. [18] using different tools, but without a quantitative lower bound on the smallest singular value.

our lower bound on how close any k columns can be to being linearly dependent in the langugage of test functions, which we use to show that when the scaling parameters are different, we can isolate one component at a time and subtract it off from the rest of the mixture. Combining these tools, we obtain our main algorithm. We note that the separation conditions we impose between pairs of components are information-theoretically necessary for our learning task.

It is natural to ask whether the dependence on k can be improved. First, we show lower bounds on the sample complexity. We construct two mixtures M and M' whose components are far apart — every pair of components has total variation distance at least μ — but M and M' have total variation distance about μ^{2k-1}. As a corollary we have:

Corollary 1. *Any algorithm for learning the components of a mixture of k Mallows models within μ in total variation distance must take at least $(1/\mu)^{2k-1}$ samples.*

Second, we consider a restricted model where the learner can only make local queries of the form: Given elements $x_1, \cdots, x_c$ and locations $i_1, \cdots, i_c$ and a tolerance τ, what is the probability that the mixture assigns x_j to location i_j for all j from 1 to c, up to an additive τ? We show that our algorithms can be implemented in the local model. Moreover, in this model, we can prove lower bounds on the dependence on n and k. We show:

Theorem I.3 (Informal)**.** *Any algorithm for learning a mixture of k Mallows models through local queries must make at least $n^{\log k}$ queries or make a query with $\tau \leq n^{-\frac{1}{2}\log k}$.*

This is reminiscent of statistical query lower bounds for other unsupervised learning problems, most notably learning mixtures of Gaussians [21]. However it is not clear how to prove lower bounds on the statistical query dimension [22], because of the complicated ways that the locations that each element is mapped to affect one another in a Mallows model, which makes it challenging to embed small hard instances into larger ones. Due to space constraints, we defer the precise statements and proofs of our lower bounds to the full version.

Finally we turn to beyond-worst case analysis and ask whether there are natural conditions on the mixture that allow us to get algorithms whose dependence on n is a fixed polynomial rather than one whose degree depends on k. Rather than requiring the centers to be far apart, we merely require their scaling parameters to be separated from one-another. We show:

Theorem I.4. *Given samples from a mixture of k Mallows models*

$$M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$$

where $|\phi_i - \phi_j| \geq \gamma$ for all $i \neq j$, $\phi_i \leq 1 - \gamma$ for all i and $n \geq 10k$, there is an algorithm whose running time and sample complexity are

$$f(\gamma, \theta, w_{\min}, k)\textit{poly}(n, \log \frac{1}{\delta})$$

for learning each center π_i exactly and the mixing weights and scaling parameters to within an additive θ, where $f(\gamma, \theta, w_{\min}, k) = \textit{poly}(1/\gamma^{k^2}, 1/\theta^{k^2}, 1/w_{min}^{k^2})$. Moreover the algorithm succeeds with probability at least $1 - \delta$.

Our algorithm leverages many of the lower bounds on the total variation distance between mixtures of Mallows models and test functions for separating one component from the others that we have established along the way.

Further Related Work

There are other natural models for distributions on permutations such as the Bradley-Terry model [23] and the Plackett-Luce model [24], [25]. Zhao et al. [26] showed that a mixture of k Plackett-Luce models is generically identifiable provided that $k \leq \lfloor \frac{n-2}{2} \rfloor !$ and gave a generalized method of moments algorithm that they proved is *consistent* — meaning that as the number of samples goes to infinity, the algorithm recovers the true parameters. More generally, Teicher [27], [28] obtained sufficient conditions for the identifiability of finite mixtures but these conditions do not apply in our setting.

II. Preliminaries

A. Basic Notation

Let $[n] = \{1, 2, \cdots, n\}$. Given two permutations π and π' on $[n]$, let $d_{KT}(\pi, \pi')$ denote the Kendall-Tau distance, which counts the number of pairs (i, j) for which the two rankings disagree.

Definition 1. A Mallows model $M(\phi, \pi^*, n)$ defines a distribution on permutations of the set $[n]$ where the probability of generating a permutation π is equal to

$$\frac{\phi^{d_{KT}(\pi^*, \pi)}}{Z_n(\phi)}$$

and $Z_n(\phi) = \sum_\pi \phi^{d_{KT}(\pi^*, \pi)}$ be the normalizing constant, which is easy to see is independent of π^*. When the number of elements is clear from context, we will omit n and write $M(\phi, \pi^*)$.

The following is a well-known (see e.g. [5]) iterative process for generating a ranking from $M(\phi, \pi^*)$: Consider the elements in rank decreasing order, according to π. When we reach the $(i+1)^{st}$ ranked element, it is inserted into each of the $i+1$ possible positions with probabilities

$$\frac{\phi^i}{1+\phi+\cdots+\phi^i}, \cdots, \frac{1}{1+\phi+\cdots+\phi^i}$$

respectively, where the order of the probabilities go from the highest rank position it could be inserted to the lowest. When the last element is inserted, the result is a random permutation drawn from $M(\phi, \pi^*)$.

A mixture of k Mallows models is defined in the usual way: We write

$$M = w_1 M(\phi_1, \pi_1^*) + \cdots + w_k M(\phi_k, \pi_k^*)$$

where the mixing weights $w_1, w_2, \cdots, w_k$ are nonnegative and sum to one. A permutation is generated by first choosing an index (each i is chosen with probability w_i) and then drawing a sample from the corresponding Mallows model $M(\phi_i, \pi_i^*)$.

We will often work with the natural vectorizations of probability distributions:

Definition 2. If P is a distribution over permutations on $[n]$ we let $v(P)$ denote the length $n!$ vector whose entries are the probabilities of generating each possible permutation. We will abuse notation and write $v(M)$ for the vectorization of a Mallows model M.

Our algorithms and their analyses will frequently make use of the notion of restricting a permutation to a set of elements:

Definition 3. Given a permutation π on $[n]$ and a subset $S \subseteq [n]$, let $\pi|_S$ be the permutation on the elements of S induced by π.

B. Block and Orders

Our algorithms will be built on various structures we impose on permutations. The way to think about these structures is that each one gives us a statistic that we can measure: What is the probability that a permutation sampled from an unknown Mallows model has the desired structure? These act like natural moments of the distribution, that we will manipulate and use in conjunction with tensor methods to design our algorithms.

Definition 4. A block structure $\mathcal{B} = S_1, S_2, \cdots, S_j$ is an ordered collection of disjoint subsets of $[n]$. We say that a permutation π satisfies $\mathcal{B}$ as a block structure if for each i, the elements of S_i occur consecutively (i.e. in positions $a_i, a_i+1, \ldots, a_i+|S_i|-1$ for some a_i) in π and moreover the blocks occur in the order $S_1, S_2, \cdots, S_j$. Finally we let $\mathcal{S}_\mathcal{B}$ denote the set of permutations satisfying $\mathcal{B}$ as a block structure.

Definition 5. An order structure $\mathcal{O} = S_1, S_2, \cdots, S_j$ is a collection of ordered subsets of $[n]$. We say a permutation π satisfies $\mathcal{O}$ as an order structure if for each i, the elements of S_i occur in π in the same relative order as they do in S_i.

Definition 6. An ordered block structure $\mathcal{A} = S_1, S_2, \cdots, S_j$ is an ordered collection of ordered disjoint subsets of $[n]$. We say a permutation π satisfies $\mathcal{A}$ as an ordered block structure if it satisfies $S_1, S_2, \cdots, S_j$ both as a block structure and as an order structure — i.e. we forget the order within each S_i when we treat it as a block structure and we forget the order among the S_i's when we treat it as an order structure.

To help parse these definitions, we include the following example:

Example 1. *Let $n = 7$ and consider $\mathcal{A} = (1,2),(4,5,6)$. The permutation $(1,2,3,7,6,5,4)$ satisfies $\mathcal{A}$ as a block structure. The permutation $(1,3,4,2,5,6,7)$ satisfies $\mathcal{A}$ as an order structure and the permutation $(1,2,3,4,5,6,7)$ satisfies $\mathcal{A}$ as an ordered block structure.*

III. Basic Facts

Here we collect some basic facts about Mallows models, in particular a lower bound on the probability that they satisfy a given block structure if their base permutation does, relationships between the total variation distance and parameter distance, and determinantal identities for special matrices.

A. What Block Structures are Likely to be Satisfied?

In this subsection, our main result is a lower bound on the probability that a permutation drawn from a Mallows model satisfies a block structure that the underlying base permutation does. Along the way, we will also establish some useful ways to think about conditioning and projecting Mallows models in terms of tensors.

Fact 1. *The conditional distribution of a Mallows model $M(\phi, \pi^*)$ when restricted to rankings where the elements in the set S (of size j) are ranked in positions $a, a+1, \cdots, a+j-1$ and the ranking of elements in $[n] \setminus S$ is fixed is precisely $M(\phi, \pi|_S^*)$.*

Proof: It is easy to see that for any two permutations τ and τ' on $[n]$ where the elements of S are

ranked in positions $a, a+1, \cdots, a+j-1$ and agree on the rankings of elements in $[n] \setminus S$ satisfy

$$d_{KT}(\pi^*, \tau) - d_{KT}(\pi^*, \tau') = d_{KT}(\pi|^*_S, \tau|_S) - d_{KT}(\pi|^*_S, \tau'|_S)$$

Thus the ratio of their probabilities is the same as the ratio of probabilities of $\tau|_S$ and $\tau'|_S$ in $M(\phi, \pi|^*_S)$, which completes the proof. ■

Next we will describe a natural way to think about the conditional distribution on permutations that satisfy a given block structure as a tensor. Recall that the subsets of $[n]$ in a block structure are required to be disjoint.

Definition 7. Given a Mallows model $M(\phi, \pi^*)$ and a block structure $\mathcal{B} = S_1, S_2, \cdots, S_j$, we define a $|S_1|! \times |S_2|! \times \cdots \times |S_j|!$ dimensional tensor $T_{M,\mathcal{B}}$ as follows: Each entry corresponds to orderings $\pi_1, \pi_2, \cdots, \pi_j$ of $S_1, S_2, \cdots, S_j$ respectively and in it, we put the probability that a ranking drawn from M satisfies $\mathcal{B}$ and for each i, the elements in S_i occur in the order specified by π_i.

It is easy to see that $T_{M,\mathcal{B}}$ has rank one. Technically this requires the obvious generalization of Fact 1 where we condition on the elements in each S_i occurring in specified consecutive locations, and then note that these events are all disjoint.

Corollary 2.

$$T_{M,\mathcal{B}} = \mathop{\mathbf{Pr}}_{M}[\pi \in \mathcal{S}_{\mathcal{B}}] \cdot v(M(\phi, \pi|_{S_1})) \otimes \cdots \otimes v(M(\phi, \pi|_{S_j}))$$

Our next result gives a convenient lower bound for the probability that a sample satisfies a given block structure B provided that the base permutation satisfies B. We defer the proof to the full version.

Lemma 1. *For any Mallows model $M(\phi, \pi^*)$ and block structure $\mathcal{B} = S_1, S_2, \cdots, S_j$ where π^* satisfies $\mathcal{B}$ and $\ell = |S_1| + \cdots + |S_j|$, we have*

$$\mathop{\mathbf{Pr}}_{M}[\pi \in \mathcal{S}_{\mathcal{B}}] \geq \frac{1}{n^{2\ell}}$$

B. Total Variation Distance Bounds

In this subsection, we give some useful relationships between the total variation distance and the parameter distance between two Mallows models (in special cases) in terms of the distance between their base permutations and scaling parameters. We will defer the proofs to the full version. First we note that if two Mallows models have different base permutations and their scaling parameters are bounded away from one, then the distributions that they generate cannot be too close.

Claim 1. *Consider two Mallows models $M_1 = M(\phi_1, \pi_1)$ and $M_2 = M(\phi_2, \pi_2)$ where $\pi_1 \neq \pi_2$ and $\phi_1, \phi_2 \leq 1 - \epsilon$. Then $d_{TV}(M_1, M_2) \geq \frac{\epsilon}{2}$.*

Second, we give a condition under which we can conclude that two Mallows models are close in total variation distance. An analogous result is proved in [11] (see Lemma 2.6) except that here we remove the dependence on ϕ_{min}.

Lemma 2. *Consider two Mallows models $M_1 = M(\phi_1, \pi)$ and $M_2 = M(\phi_2, \pi)$ with the same base permutation on $n \geq 2$ elements. If $|\phi_1 - \phi_2| \leq \frac{\mu^2}{10n^3}$ then $d_{TV}(M_1, M_2) \leq \mu$.*

C. Special Matrix Results

Here we present a determinantal identity from mathematical physics that will play a central role in our learning algorithms. Note $d_{KT}(\pi, \sigma) = I(\pi\sigma^{-1})$ where I counts the number of inversions in a permutation. We make the following definition.

Definition 8. Let $A_n(\phi)$ be the $n! \times n!$ matrix whose rows and columns are indexed by permutations π, σ on $[n]$ and whose entries $A_{\pi\sigma}$ are $\phi^{I(\pi\sigma^{-1})}$.

Zagier [1] gives us an explicit form for the determinant of this matrix, which we quote here:

Theorem III.1. *[1]*

$$\det(A_n(\phi)) = \prod_{i=1}^{n-1} (1 - \phi^{i^2+i})^{\frac{n!(n-i)}{i^2+i}}$$

This expression for the determinant gives us weak lower bounds on the total variation distance between mixtures of Mallows models where all the scaling parameters are the same. We will bootstrap this identity to prove a stronger result about how far a column of A is from the span of any set of $k-1$ other columns.

IV. Identifiability

In this section we show that any two mixtures of k Mallows models whose components are far from each other (and the uniform distribution) in total variation distance are far from each other as mixtures too, provided that $n > 10k^2$.

A. Robust Kruskal Rank

Our first step is to show that any k columns of $A_n(\phi)$ are not too close to being linearly dependent — i.e. the projection of any column onto the orthogonal complement of the span of any $k-1$ other columns cannot be too small. The *Kruskal rank* of a collection of vectors is the largest ℓ so that every ℓ vectors are

linearly independent. The property we establish here is sometimes called a *robust Kruskal rank* [20].

Lemma 3. *Suppose $\phi < 1-\epsilon$ and consider k columns of $A_n(\phi)$. The projection of one column onto the orthogonal complement of the other $k-1$ has euclidean length at least $(\frac{\epsilon^n}{\sqrt{n!}})^k$.*

Proof: Assume for the sake of contradiction that there is a set of k columns that violates the statement of the lemma. In particular suppose that the projection of c_k onto the orthogonal complement of $c_1, c_2, \cdots, c_{k-1}$ has euclidean length N for some $N < (\frac{\epsilon^n}{\sqrt{n!}})^k$. We will use this assumption to prove an upper bound on the determinant of $A_n(\phi)$ that violates Theorem III.1. Our approach is to find an ordering of the columns so that as we scan through, at least once every k columns the euclidean length of its projection onto the orthogonal complement of the columns seen so far is at most N. Then using the naive upper bound of $\sqrt{n!}$ on the euclidean length of any column of $A_n(\phi)$ we have

$$\det(A_n(\phi)) \leq N^{\frac{n!}{k}} (\sqrt{n!})^{n!}$$

However from Theorem III.1 we have

$$\begin{aligned}\det(A_n(\phi)) &= \prod_{i=1}^{n-1} (1-\phi^{i^2+i})^{\frac{n!(n-i)}{i^2+i}} \\ &\geq (1-\phi)^{n!n(\sum_{i=1}^{n-1} \frac{1}{i(i+1)})} \geq (1-\phi)^{n!n} \\ &\geq \epsilon^{n!n}\end{aligned}$$

which yields the desired contradiction.

Now we complete the argument by constructing the desired ordering of the columns as follows: We start with $c_1, c_2, \cdots, c_k$. And then we choose any column c not yet selected. Let π be the permutation that maps c_k to c. Now π maps $c_1, c_2, \cdots, c_{k-1}$ to $k-1$ columns, and suppose j of them have not been selected yet. Call these $c'_1, c'_2, \cdots, c'_j$. We continue the ordering of the columns by appending $c'_1, c'_2, \cdots, c'_j, c$. It is easy to see that the euclidean length of the projection of c onto the orthogonal complement of the columns seen so far is also at most N, which now finishes the proof. ∎

The above lemma is not directly useful for two reasons: First, the lower bound is exponentially small in n. Second, it is tantamount to a lower bound on the ℓ_2-norm of any sparse linear combination of the columns of $A_n(\phi)$. What we really want in the context of identifiability is a lower bound on the ℓ_1-norm (of a matrix whose columns represent the components).

Definition 9. Let $B_n(\phi)$ be obtained from $A_n(\phi)$ by normalizing its columns to sum to one.

Lemma 4. *Suppose $\phi < 1-\epsilon$ and consider any k columns $c_1, c_2, \ldots, c_k$ of $B_n(\phi)$. Then*

$$\|z_1 c_1 + \cdots + z_k c_k\|_1 \geq \frac{1}{n^{4k}} \frac{\epsilon^{2k^2}}{(k+1)^{k^2+2k}}$$

provided that $\max(|z_1|, |z_2|, \ldots, |z_k|) \geq 1$.

Proof: Let $\pi_1, \pi_2, \cdots, \pi_k$ be the permutations corresponding to the columns $c_1, c_2, \ldots, c_k$. Also without loss of generality suppose $z_1 \geq 1$ and that $\pi_1 = (1, 2, \cdots, n)$. First we build a block structure that π_1 satisfies but no other π_i does: For each i, pick two consecutive elements in π_1, say x_i and $x_i + 1$ that are inverted in π_i. Such a pair exists because $\pi_1 \neq \pi_i$. Now we can take the union of these pairs over all i to form a block structure $\mathcal{B} = \{S_1, S_2, \cdots, S_j\}$ so that, for all i, x_i and $x_i + 1$ are in the same block and π_1 satisfies $\mathcal{B}$. Note that j can be less than k, if for example two of the pairs contain the same element. In any case, we have $|S_1| + \cdots + |S_j| \leq 2k$.

Now for each i, set $M_i = M(\phi, \pi_i)$ and $T_i = T_{M_i, \mathcal{B}}$. From Corollary 2 we have that

$$T_i = \mathop{\mathbf{Pr}}_{M_i}[\pi \in \mathcal{S}_{\mathcal{B}}] \cdot v(M(\phi, \pi_i|_{S_1})) \otimes \cdots \otimes v(M(\phi, \pi_i|_{S_j}))$$

Next we show that we can find unit vectors $v_1, \cdots, v_j$ so that

(1) $\langle v_b, v(M(\phi, \pi_i|_{S_b}))\rangle = 0$ whenever $\pi_i|_{S_b} \neq \pi_1|_{S_b}$ and

(2) $\langle v_b, v(M(\phi, \pi_1|_{S_b}))\rangle \geq \frac{1}{|S_b|!}(\frac{\epsilon^{|S_b|}}{\sqrt{|S_b|!}})^k$ for all b

This fact essentially follows from Lemma 3. First observe that the $v(M(\phi, \pi_i|_{S_b}))$ and the column of $A_{|S_b|}(\phi)$ corresponding to $\pi_i|_{S_b}$ differ only by a normalization, since the former sums to one. Now for each b we can take v_b to be the unit vector in the direction of the projection of $v(M(\phi, \pi_1|_{S_b}))$ onto the orthogonal complement of all the $v(M(\phi, \pi_i|_{S_b}))$'s for $i \neq 1$. Note that the additional $\frac{1}{|S_b|!}$ factor arises because Lemma 3 deals with $A_{|S_b|}(\phi)$ and to normalize any column of it we need to divide by at most $|S_b|!$.

With this construction, we have that $\langle v_1 \otimes v_2 \otimes \cdots \otimes v_j, T_i\rangle = 0$ for all $i \neq 1$ since each π_i differs from π_1 when restricted to at least one of the blocks $S_1, S_2, \cdots, S_j$. Moreover using property (2) above and Lemma 1 we have

$$\begin{aligned}\langle v_1 \otimes v_2 \otimes \cdots \otimes v_j, T_1\rangle &\geq \frac{1}{n^{4k}} \prod_b \frac{1}{|S_b|!} (\frac{\epsilon^{|S_b|}}{\sqrt{|S_b|!}})^k \\ &\geq \frac{1}{n^{4k}} \frac{1}{(2k)!} (\frac{\epsilon^{2k}}{\sqrt{(2k)!}})^k \geq \frac{1}{n^{4k}} \frac{\epsilon^{2k^2}}{(k+1)^{k^2+2k}}\end{aligned}$$

where the last inequality follows from the bound $(2k)! \leq (k+1)^{2k}$. Finally note that

$$\|z_1 c_1 + \cdots + z_k c_k\|_1 \geq \sum_i \langle v_1 \otimes v_2 \otimes \cdots \otimes v_j, T_i \rangle$$

since the entries of $v_1 \otimes v_2 \otimes \cdots \otimes v_j$ are at most one in absolute value, and each T_i can be formed from c_i by zeroing out entries (corresponding to permutations that do not satisfy $\mathcal{S}_{\mathcal{B}}$) and summing subsets of the remaining ones together (that correspond to permutations with the same ordering for each S_i). ■

The above lemma readily implies that any two mixtures of k Mallows models whose components all have the same scaling parameter and whose mixing weights are different are far from each other in total variation distance. In the sequel we will be interested in proving identifiability even when the scaling parameters are allowed to be different. As a step towards that goal, first we give a simple extension that allows the scaling parameters to be slightly different. We defer the proof of the following lemma to the full version.

Lemma 5. *Consider any k distinct permutations $\pi_1, \pi_2, \cdots, \pi_k$ and scaling parameters $\phi_1, \phi_2, \cdots, \phi_k$. Let $c_i = v(M(\phi_i, \pi_i))$ and suppose that for each i, $\phi_i \leq 1 - \epsilon$ and for each $i \neq j$,*

$$|\phi_i - \phi_j| \leq \frac{1}{160 n^{8k+3}} \frac{\epsilon^{4k^2}}{(k+1)^{2k^2+4k+2}}$$

Then for any coefficients z_i with $\max(|z_1|, |z_2|, \ldots, |z_k|) \geq 1$ we have

$$\|z_1 c_1 + \cdots + z_k c_k\|_1 \geq \frac{1}{2n^{4k}} \frac{\epsilon^{2k^2}}{(k+1)^{k^2+2k}}$$

B. Polynomial Identifiability

Now we are ready to prove our main identifiability result. First we define a notion of non-degeneracy, which is information-theoretically necessary when our goal is to identify all the components in the mixture.

Definition 10. We say a mixture of Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$ is μ-non degenerate if the total variation distance between any pair of components is at least μ and the total variation distance between any component and the uniform distribution over all permutations is also at least μ. Furthermore we say that the mixture is (μ, α)-non degenerate if in addition each mixing weight is at least α.

We will not need the following definition until later (when we state the guarantees of various intermediary algorithms), but let us also define a natural notion for two mixtures to be component-wise close:

Definition 11. We say that two mixtures of Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$ and $M' = w'_1 M(\phi'_1, \pi'_1) + \cdots + w'_k M(\phi'_k, \pi'_k)$ with the same number of components are component-wise θ-close if there is a relabelling of components in one of the mixtures after which $|w_i - w'_i|, |\phi_i - \phi'_i|$ and $d_{TV}(M(\phi_i, \pi_i), M(\phi'_i, \pi'_i)) \leq \theta$ for all i.

Most of these conditions are standard, because in order to be able to identify M from a polynomial number of samples we need to get at least one sample from each component and at least one sample from the difference between any two components. In our context, we additionally require no component to be too close to the uniform distribution because the distribution $M(1, \pi)$ is the same regardless of the choice of π.

Below, we state our main lemma in this section. The technical argument is quite involved and we defer the details of the proof to the full version, but many of the antecedents (in particular finding block structures that capture the disagreements between permutations, representing the distribution on a subset of permutations as a tensor and constructing test functions as the tensor product of simple vectors) were used already in the proof of Lemma 4.

Lemma 6. *Consider any k (not necessarily distinct) permutations $\pi_1, \pi_2, \cdots, \pi_k$ and scaling parameters $\phi_1, \phi_2, \cdots, \phi_k$. Set $M_i = M(\phi_i, \pi_i)$ and suppose that the collection of Mallows models is μ-non degenerate. Then for any coefficients z_i with $\max(|z_1|, |z_2|, \ldots, |z_k|) \geq 1$ we have*

$$\|z_1 v(M_1) + \cdots + z_k v(M_k)\|_1 \geq \Big(\frac{\mu^2}{10 n^4 k}\Big)^{20k^3}$$

Now we present our main identifiability result. We defer the proof to the full version as it is essentially an application of the previous lemma.

Theorem IV.1. *Consider two mixtures of Mallows models*

$$M = w_1 M(\phi_1, \pi_1) + \cdots + w_i M(\phi_i, \pi_i) \text{ and}$$
$$M' = w'_1 M(\phi'_1, \pi'_1) + \cdots + w'_{i'} M(\phi'_{i'}, \pi_{i'})$$

where $i, i' \leq k$. Suppose that both mixtures are (μ, α)-non degenerate and set $\epsilon = \frac{\mu^2}{10 n^3}$. If for some parameter θ we have

$$\|\sum_i w_i v(M(\phi_i, \pi_i)) - \sum_{i'} w'_{i'} v(M(\phi_{i'}, \pi_{i'}))\|_1 \leq \Big(\frac{\epsilon \theta \alpha}{nk}\Big)^{(10k)^{6k}}$$

Then $i = i'$ and there is a matching between the components in the two mixtures so that across the matching the components are θ-close in total variation distance and the mixing weights are also θ-close.

V. The General Algorithm

Here we leverage the tools and ideas that we developed in the previous section to give a polynomial time algorithm for learning mixtures of Mallows models that works for any constant number of components. We have already seen the key ingredient — test functions that isolate a single component from the rest of the mixture. In the context of polynomial identifiability, we knew the parameters of the mixture which we used to construct small ordered block structures that allow us to focus on parts of the distribution that have a convenient analytic form, but still capture the differences between the base permutations. Here we will use variants of the same type of arguments, but where we guess the relevant portions of the ordered block structures from which we follow the same recipe to construct test functions. If our guess is correct, we will succeed in learning the base permutation of some component. In our setting there will be a constant number of guesses, so we will be able to construct a list of candidate mixtures at least one of which is close to the true mixture. We can then appeal to our identifiability results to test find a mixture that is close on a component-wise basis.

Because we will need to handle components with small scaling parameters separately, it will be more convenient to work with vectorizations of the low order moments of a distribution than the distribution itself.

Definition 12. For a Mallows model M on n elements, let $v_c(M)$ denote the vectorization of the order c moments of M. In particular $v_c(M)$ has $\binom{n}{c}n(n-1)\cdots(n-c+1)$ entries and we interpret each coordinate as a choice of a subset $S \subset [n]$ of size c and a placement of its elements. The value of the entry in $v_c(M)$ is the probability under M that the elements in S are placed in the corresponding locations.

Note that the sum of the entries in $v_c(M)$ is larger than one, because the values of its entries are the probabilities of events that are (usually) not disjoint. We remark that all of the proofs of polynomial identifiability, where we prove lower bounds on the ℓ_1-norm of linear combinations of vectorizations of Mallows models, carry over to the case when we use $v_c(M)$ instead provided that $c \geq 10k^2$. This follows by observing that all of the events we used can be defined in terms of the placement of at most c elements and the probabilities of these events can be computed by adding up an appropriate set of the entries of $v_c(M)$ (corresponding to disjoint events themselves) instead of $v(M)$. Thus we have the following corollary:

Corollary 3. *Consider two mixtures of Mallows models*

$$M = w_1 M(\phi_1, \pi_1) + \cdots + w_i M(\phi_i, \pi_i) \text{ and}$$
$$M' = w'_1 M(\phi'_1, \pi'_1) + \cdots + w'_{i'} M(\phi'_{i'}, \pi_{i'})$$

where $i, i' \leq k$. Suppose that both mixtures are (μ, α)-non degenerate and set $\epsilon = \frac{\mu^2}{10n^3}$. If for some parameter θ we have

$$\|\sum_i w_i v_c(M(\phi_i, \pi_i)) - \sum_{i'} w'_{i'} v_c(M(\phi_{i'}, \pi_{i'}))\|_1 \leq \left(\frac{\epsilon\theta\alpha}{nk}\right)^{(10k)^{6k}}$$

Then $i = i'$ and there is a matching between the components in the two mixtures so that across the matching the components are θ-close in total variation distance and the mixing weights are also θ-close.

Next we give an outline of our algorithm: In Claim 2 we give an algorithm for finding and removing components with small scaling parameter. The intuition is that such components often generate their own base permutation, so if we take a small number of samples and find all the permutations that occur somewhat frequently, we will have a superset of the base permutations of components with small scaling parameter. We then remove their contribution to the order c moments to generate a list of candidate vectors, at least one of which is close to the true order c moments of the submixture of components without small scaling parameter. In Corollary 4 we give an algorithm that mimics the proof of Lemma 4 and Lemma 6 but wherever the construction of a test function to isolate a component uses knowledge of the mixture, we guess. The algorithm outputs a list of candidate parameters with the property that accurate estimates of each component in the true mixture appear on the list. We then consider all k tuples of components to form a list of candidate mixtures. Finally in Corollary 5 we give an algorithm for testing whether a candidate mixture is close to the true mixture. The intuition is we can generate our own samples from a candidate mixture to compute its lower order moments and check whether these are close to the lower order moments of the true mixture. Corollary 3 tells us that if this check passes then the candidate mixture is indeed component-wise close to the true mixture.

A. Finding Components with Small Scaling Parameters

When the scaling parameter of a Mallows model is small enough, it generates its own base permutation

the majority of the time. Using this intuition, we show that we can take few samples and guess the base permutations of all the components with small scaling parameter and then essentially remove them from the mixture. We defer the proof of the following claim to the full version.

Claim 2. *Consider a mixture of Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_i M(\phi_i, \pi_i)$ where all mixing weights are at least α. Suppose that $\phi_1, \ldots, \phi_j < \frac{1}{2n}$ and $\phi_{j+1}, \ldots, \phi_k \geq \frac{1}{2n}$. Let $c = 10k^2$. There is an algorithm that takes*

$$m = \Big(\frac{nk \log \frac{1}{\delta}}{\epsilon\theta\alpha}\Big)^{(10k)^{8k}}$$

samples and runs in polynomial time and outputs a list of vectors v' of polynomial size so that with probability at least $1 - \frac{\delta}{3}$ at least one v' satisfies

$$\|v' - \sum_{a=j+1}^{k} w_a v_c(M_a)\|_1 \leq \frac{1}{k}\Big(\frac{\epsilon\alpha}{2nk}\Big)^{60k^4}$$

B. Finding a Single Component

In this subsection, we focus on the problem of recovering a single component from a mixture of Mallows models. Our algorithms will closely parallel the polynomial identifiability results in Lemma 4 and Lemma 6. More precisely, we will modify the test functions we used in those results to turn them into algorithms for isolating a single component, first focusing on the case when all the scaling parameters are the same and then the general setting where they can be different. We defer the proofs of the results in this section to the full version.

Lemma 7. *Consider a collection of k Mallows models $M(\phi, \pi_1), \cdots, M(\phi, \pi_k)$ with distinct base permutations and where $\phi \leq 1 - \epsilon$. Suppose ϕ is known and we are given a vector v with*

$$\|v - \sum_i z_i v_c(M(\phi, \pi_i))\|_1 \leq \Big(\frac{\epsilon^k}{nk}\Big)^{4k}$$

and a constant $c \geq 10k^2$. Finally suppose $z_1 \geq 1$. There is a polynomial time algorithm to output a list of at most $n^{4k}((2k)!)^k$ permutations that contains π_1.

Next we give an algorithm for isolating a single component when the scaling parameters are allowed to be different. In addition to the usual assumptions, we will assume that each scaling parameter is at least $\frac{1}{2n}$, since the algorithm in Claim 2 allows us to remove such components from our estimates of the order c moments.

Lemma 8. *Consider a mixture of k Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$ where for each i, $\alpha \leq w_i$ and the total variation distance between any two components is at least μ. Furthermore suppose that for all i, $\frac{1}{2n} < \phi_i < 1 - \epsilon$ where $\epsilon = \frac{\mu^2}{10n^3}$. Let $c = 10k^2$ and θ be the target accuracy. Suppose we are given a vector v with*

$$\|v - \sum_i w_i v_c(M(\phi_i, \pi_i))\|_1 \leq \Big(\frac{\epsilon\alpha}{2nk}\Big)^{60k^4}$$

There is a polynomial time algorithm to output a list of candidate parameters (w, ϕ, π) so that for some ℓ there is at one entry in the list that is θ-close — i.e. it satisfies $|w - w_\ell| \leq \theta, |\phi - \phi_\ell| \leq \theta$ and $\pi = \pi_\ell$.

C. Finding the Rest of the Components

It is now straightforward to repeatedly use the algorithm in Lemma 8 to learn and peel off components one by one, which is captured by the following corollary:

Corollary 4. *Under the same conditions as Lemma 8, suppose we are given a vector v with*

$$\|v - \sum_i w_i v_c(M(\phi_i, \pi_i))\|_1 \leq \Big(\frac{\epsilon\alpha}{2nk}\Big)^{60k^4}$$

There is a polynomial time algorithm to output a list of candidate parameters (w, ϕ, π) so that for each ℓ there is at one entry in the list that is θ-close — i.e. it satisfies $|w - w_\ell| \leq \theta, |\phi - \phi_\ell| \leq \theta$ and $\pi = \pi_\ell$.

Of course, once we have a list where an accurate estimate of every component in the mixture appears, we can try all possible k tuples of parameters on the list in order to generate a list of candidate mixtures, at least one of which is component-wise close to the true mixture. All that remains is to hypothesis test all of these possibilities. This is slightly more complicated in our setting, because we want to select a candidate that is not just close in total variation distance as a mixture, but even in a component-wise sense.

D. Testing Component-wise Closeness

Here we show how to test whether a pair of mixtures of Mallows models are component-wise close. This essentially follows by invoking Corollary 3 and standard arguments.

Corollary 5. *Suppose we are given sample access to a mixture of k Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$ and an estimate $M' = w'_1 M(\phi'_1, \pi'_1) + \cdots + w_k M(\phi'_k, \pi'_k)$ on $n \geq 10k^2$ elements where both*

mixtures are (μ, α)-non degenerate. Se t $\epsilon = \frac{\mu^2}{10n^3}$. There is an algorithm which given

$$m = \Big(\frac{nk \log \frac{1}{\delta'}}{\epsilon\theta\alpha}\Big)^{(10k)^{8k}}$$

samples from M runs in polynomial time and succeeds in accepting when the mixtures are component-wise γ-close for

$$\gamma = \Big(\frac{\epsilon\theta\alpha}{nk \log \frac{1}{\delta'}}\Big)^{(10k)^{10k}}$$

and rejects when they are component-wise θ-far and succeeds with probability at least $1 - \delta'$.

Proof: As usual, set $c = 10k^2$. From the choice of γ we have that if the mixtures are component-wise γ-close then

$$\|\sum_i w_i v_c(M(\phi_i, \pi_i)) - \sum_i w'_i v_c(M(\phi'_i, \pi'_i))\|_1 \leq \Big(\frac{\epsilon\theta\alpha}{nk}\Big)^{(10k)^{8k}}$$

And from Corollary 3 we know a weak converse that if the above bound (with, say, an extra factor of four) holds then the mixtures must be component-wise θ-close. Now we can estimate the above quantity using m samples from M and by generating m samples from M'. The latter can be done efficiently through any of the known sampling schemes for Mallows models, e.g. [5]. We accept if the above bound holds (with an extra factor of two) and reject otherwise, and by standard concentration argument it is easy to see that the failure probability is at most δ'. ∎

We can now complete the proof of Theorem I.1 by following the outline at the beginning of Section V. We defer the details to the full version.

VI. Beyond Worst-Case Analysis

Motivated by our lower bound against local algorithms, it is natural to ask whether there is some notion of beyond worst-case analysis whereby we can get much faster algorithms that work under tame conditions on the input. Here we give such an algorithm in the case when all of the scaling parameters are separated from each other (and from the value one, which causes a different sort of degeneracy).

Definition 13. We say a mixture of Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$ is (γ, α)-separated if $w_i \geq \alpha$ and $\phi_i \leq 1 - \gamma$ for all i and additionally $|\phi_i - \phi_j| \geq \gamma$.

In our main lemmas, we will also need some new notions of component-wise closeness that will arise due to some subtleties at intermediate steps.

Definition 14. We say that two mixtures of Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$ and $M' = w'_1 M(\phi'_1, \pi'_1) + \cdots + w'_k M(\phi'_k, \pi'_k)$ with the same number of components are component-wise θ-close in parameters if there is a relabelling of components in one of the mixtures after which $|w_i - w'_i|, |\phi_i - \phi'_i| \leq \theta$ and $\pi_i = \pi'_i$ for all i. If all but the condition on the mixing weights holds, we will say that they are component-wise θ-close in base parameters.

Theorem VI.1. *Consider a mixture of k Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$ that is (γ, α)-separated and has $n \geq 10k$. There is an algorithm that runs in time polynomial in n, $\log 1/\delta$ and $(\frac{1}{\theta\gamma\alpha})^{k^2}$ time and outputs a mixture M' that is component-wise θ-close in parameters to M with probability at least $1 - \delta$.*

We give a brief outline of our algorithm. First we run the algorithm in Claim 3 to find candidate prefixes for all the base permutations and estimates of their scaling parameters. Next we run the algorithm in Lemma 9 to extend the prefixes to a list of candidate full base permutations. Then we run the algorithm in Claim 4 so that for every k tuple of candidate base permutations and scaling parameters, we can uniquely determine accurate estimates of the mixing weights. Finally we run the algorithm in Lemma 10 to test whether a given mixture is component-wise close in parameters to the true mixture.

A. Finding the Prefixes

Our first step is to determine the first $10k$ elements of each base permutation. More precisely, given samples from M, we want to find a list of candidate prefixes, so that for each base permutation its prefix appears on the list. Along the way, we will also perform a grid search over the scaling parameters. We defer the proof of the following claim to the full version.

Claim 3. *Consider a mixture of k Mallows models $M = w_1 M(\phi_1, \pi_1) + \cdots + w_k M(\phi_k, \pi_k)$ that is (γ, α)-separated. There is an algorithm that takes*

$$m = \Big(\frac{1}{2k\alpha\gamma^{10k} \log \frac{1}{\delta}}\Big)^2$$

samples and runs in time polynomial in n, $1/(\alpha\beta)^k$, $1/\gamma^{k^2}$ and $\log 1/\delta$ and outputs a list of size

$$s = 2^k \frac{1}{\gamma^{10k^2}} \frac{1}{(\alpha\beta)^k}$$

where each entry is a candidate prefix of $10k$ elements and a scaling parameter. Moreover with probability at

least $1-\delta$, for each i, the first $10k$ elements of π_i appear as an entry in the list along with an estimate of ϕ_i that is close within an additive β.

This is the only step in our algorithm which is not obvious how to implement using local queries. Nevertheless, it can be: First, query the probability that each of the n elements appears as the first element in a draw from M. Now for each of the elements that occurs first with probability at least $\frac{1}{2}\gamma^{10k}\alpha$, query the probability of all possible second elements (conditioned on the choice of the first element). We can repeat this process to find the heavy hitters among all possible prefixes of $10k$ elements, and can remove any prefix that does not occur with probability at least $\frac{1}{2}\gamma^{10k}\alpha$ thus ensuring that the list of queries we need to make does not ever become too large.

B. Finding the Full Permutations and Mixing Weights

In this subsection, we show how to find all the base permutations from their prefixes. We will also show how to recover the mixing weights using the base permutations and scaling parameters. We defer the proofs of the results in this section to the full version.

Lemma 9. *Suppose the conditions of Theorem VI.1 hold. Suppose that $n \geq 10k$ and the first $10k$ elements of each permutation are known and we are given estimates ϕ_i' of the scaling parameters that, for each i, satisfy $|\phi_i - \phi_i'| \leq \beta$ with $\beta \leq \alpha\left(\frac{\gamma}{10}\right)^{3k}$. There is an algorithm that runs in time polynomial in n, $1/\alpha$, $1/\beta$, $\log 1/\delta$ and $1/\gamma^k$, uses $m = \left(2^k n \frac{1}{\beta}\log\frac{1}{\delta}\right)^2$ samples, and outputs a list of 2^{k-1} permutations for each i so that with probability at least $1-\delta$, each π_i is included on the corresponding list.*

It might seem like we can now just grid search over the mixing weights. But there is a subtle issue: If all the base permutations and scaling parameters are correct, but the mixing weights are not, our testing algorithm might not be able to tell. For this reason we need to make sure that once we have a candidate set of base permutations and their scaling parameters, we do not need to do any guessing to determine the mixing weights. The following claim shows how the ideas in the proof of Lemma 9 can be adapted to resolve this issue.

Claim 4. *Suppose the conditions of Theorem VI.1 hold. Furthermore suppose the base permutations are known and we are given estimates ϕ_i' of the scaling parameters that, for each i, satisfy $|\phi_i - \phi_i'| < \beta$ with $\beta \leq \alpha^2\left(\frac{\gamma}{10}\right)^{6k}$. There is an algorithm that runs in time polynomial in n, $1/\beta$ and $\log 1/\delta$ time and uses $m = \left(2^k n \frac{1}{\beta}\log\frac{1}{\delta}\right)^2$ samples and outputs estimates of the mixing weights that satisfy*

$$|w_i - w_i'| \leq \frac{\beta}{\alpha}\left(\frac{10}{\gamma}\right)^{4k}$$

for each i, with probability at least $1-\delta$.

C. Testing Closeness for Separated Mixtures

The final piece of our algorithm is a method for testing if two $(\frac{\gamma}{2},\alpha)$-separated mixtures M and M' are component-wise close in parameters. We defer the proof to the full version.

Lemma 10. *Suppose we are given sample access to a mixture of k Mallows models $M = w_1 M(\phi_1,\pi_1)+\cdots+ w_k M(\phi_k,\pi_k)$ and an estimate $M' = w_1' M(\phi_1',\pi_1') + \cdots + w_k M(\phi_k',\pi_k')$ and both are $(\frac{\gamma}{2},\alpha)$-separated. Finally suppose $n \geq 10k$ and $\theta \leq \frac{\gamma}{10}$. There is an algorithm which given*

$$m = \left(\frac{nk10^k \log\frac{1}{\delta}}{\theta\alpha\gamma^k}\right)^{20}$$

samples from M runs in polynomial in n, $1/\gamma^k$, $1/\alpha$, $1/\theta$ and $\log 1/\delta$ time and if M and M' are not component-wise θ-close in base parameters, rejects with probability at least $1-\delta$. And if they are component wise θ'-close in parameters, for

$$\theta' = \left(\frac{\theta\alpha\gamma^k}{10^k}\right)^{50}$$

it accepts with probability at least $1-\delta$.

We can now complete the proof of Theorem I.4 by following the outline at the beginning of Section VI. We defer the details to the full version.

ACKNOWLEDGMENT

AM was supported in part by NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, a David and Lucile Packard Fellowship and an Alfred P. Sloan Fellowship.

REFERENCES

[1] D. Zagier, "Realizability of a model in infinite statistics," *Comm. Math. Phys.*, vol. 147, no. 1, pp. 199–210, 1992. [Online]. Available: https://projecteuclid.org:443/euclid.cmp/1104250533

[2] H. P. Young, "Condorcet's theory of voting," *American Political science review*, vol. 82, no. 4, pp. 1231–1244, 1988.

[3] C. J. Burges, "From ranknet to lambdarank to lambdamart: An overview," *Learning*, vol. 11, no. 23-581, p. 81, 2010.

[4] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938. [Online]. Available: +http://dx.doi.org/10.1093/biomet/30.1-2.81

[5] J.-P. Doignon, A. Pekeč, and M. Regenwetter, "The repeated insertion model for rankings: Missing link between two subset choice models," *Psychometrika*, vol. 69, no. 1, pp. 33–54, 2004.

[6] I. C. Gormley and T. B. Murphy, "Exploring voting blocs within the irish electorate: A mixture modeling approach," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1014–1027, 2008.

[7] T. Lu and C. Boutilier, "Effective sampling and learning for mallows models with pairwise-preference data," *Journal of Machine Learning Research*, vol. 15, pp. 3963–4009, 2014. [Online]. Available: http://jmlr.org/papers/v15/lu14a.html

[8] T. B. Murphy and D. Martin, "Mixtures of distance-based models for ranking data," *Computational statistics & data analysis*, vol. 41, no. 3-4, pp. 645–655, 2003.

[9] M. Meila and H. Chen, "Dirichlet process mixtures of generalized mallows models," in *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, 2010, pp. 358–367.

[10] M. Braverman and E. Mossel, "Noisy sorting without resampling," in *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2008, pp. 268–276.

[11] P. Awasthi, A. Blum, O. Sheffet, and A. Vijayaraghavan, "Learning mixtures of ranking models," in *Advances in Neural Information Processing Systems*, 2014, pp. 2609–2617.

[12] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *arXiv preprint arXiv:1210.7559*, 2012.

[13] Y. Freund and Y. Mansour, "Estimating a mixture of two product distributions," in *Proceedings of the twelfth annual conference on Computational learning theory*. ACM, 1999, pp. 53–62.

[14] J. Feldman, R. ODonnell, and R. A. Servedio, "Learning mixtures of product distributions over discrete domains," in *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 501–510. [Online]. Available: http://dx.doi.org/10.1109/SFCS.2005.46

[15] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently learning mixtures of two gaussians," in *Proceedings of the 42nd ACM symposium on Theory of computing*. ACM, 2010, pp. 553–562.

[16] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of gaussians," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE, 2010, pp. 93–102.

[17] M. Belkin and K. Sinha, "Polynomial learning of distribution families," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE, 2010, pp. 103–112.

[18] F. Chierichetti, A. Dasgupta, R. Kumar, and S. Lattanzi, "On learning mixture models for permutations," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 2015, pp. 85–92.

[19] N. Bhatnagar and R. Peled, "Lengths of monotone subsequences in a mallows permutation," *Probability Theory and Related Fields*, vol. 161, no. 3-4, pp. 719–780, 2015.

[20] E. S. Allman, C. Matias, and J. A. Rhodes, "Identifiability of parameters in latent structure models with many observed variables," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3099–3132, 2009.

[21] I. Diakonikolas, D. M. Kane, and A. Stewart, "Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures," in *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*. IEEE, 2017, pp. 73–84.

[22] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao, "Statistical algorithms and a lower bound for detecting planted cliques," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 655–664.

[23] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[24] R. L. Plackett, "The analysis of permutations," *Applied Statistics*, pp. 193–202, 1975.

[25] R. D. Luce, *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

[26] Z. Zhao, P. Piech, and L. Xia, "Learning mixtures of plackett-luce models," in *International Conference on Machine Learning*, 2016, pp. 2906–2914.

[27] H. Teicher, "Identifiability of mixtures," *The annals of Mathematical statistics*, vol. 32, no. 1, pp. 244–248, 1961.

[28] ——, "Identifiability of finite mixtures," *The annals of Mathematical statistics*, pp. 1265–1269, 1963.