Identification of Mixtures of Discrete Product Distributions in Near-Optimal Sample and Time Complexity

Spencer L. Gordon SLGORDON@CALTECH.EDU
Erik Jahn EJAHN@CALTECH.EDU
Engineering and Applied Science, California Institute of Technology, Pasadena CA 91125, USA.

Bijan Mazaheri BMAZAHER@BROADINSTITUTE.ORG
Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard, Cambridge MA 02142, USA.

Yuval Rabani YRABANI@CS.HUJI.AC.IL
The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 9190416, Israel.

Leonard J. Schulman SCHULMAN@CALTECH.EDU
Engineering and Applied Science, California Institute of Technology, Pasadena CA 91125, USA. *

Editors: Shipra Agrawal and Aaron Roth

Abstract

We consider the problem of *identifying,* from statistics, a distribution of discrete random variables $X_1, \ldots, X_n$ that is a mixture of k product distributions. The best previous sample complexity for $n \in O(k)$ was $(1/\zeta)^{O(k^2 \log k)}$ (under a mild separation assumption parameterized by ζ). The best known lower bound was $\exp(\Omega(k))$.

It is known that $n \geq 2k-1$ is necessary and sufficient for identification. We show, for any $n \geq 2k-1$, how to achieve sample complexity and run-time complexity $(1/\zeta)^{O(k)}$. We also extend the known lower bound of $e^{\Omega(k)}$ to match our upper bound across a broad range of ζ.

Our results are obtained by combining (a) a classic method for robust tensor decomposition, (b) a novel way of bounding the condition number of key matrices called Hadamard extensions, by studying their action only on flattened rank-1 tensors.

1. Introduction

1.1. The problem and our results.

This paper resolves the sample and runtime complexity of identification of mixtures of product distributions, a problem introduced thirty years ago in Kearns et al. (1994), and further studied in Cryan et al. (2001); Freund and Mansour (1999); Hall and Zhou (2003); Feldman et al. (2008); Chaudhuri and Rao (2008); Tahmasebi et al. (2018); Chen and Moitra (2019); Gordon et al. (2021). In this problem, an observer collects samples from a distribution over n binary (or otherwise drawn from a small finite set) random variables $X_1, X_2, \ldots, X_n$. The samples are collected from a mixture of k distinct sub-populations. For each sample of $(X_1, \ldots, X_n)$, a sub-population $U \in \{1, \ldots, k\}$ is chosen independently from previous samples and according to the frequencies of the sub-populations in the entire population. Then, the random variables X_i are drawn independently conditional on the sub-population U. The observer does not know the frequencies of the sub-populations, and does not get an indication of the sub-population from which a sample was drawn; the observer only sees the values of the n observable random bits. The goal is to reconstruct the probabilistic model that generated the collected samples, namely, the frequencies of the sub-populations and the conditional product distributions on $(X_1, X_2, \ldots, X_n) \in \{0,1\}^n$.

* Research supported by NSF CCF-1909972 and CCF-2321079, by ISF grants 3565-21 and 389-22, and by BSF grant 2023607.

© 2024 S.L. Gordon, E. Jahn, B. Mazaheri, Y. Rabani & L.J. Schulman.

Most of the above literature discusses the problem of *learning* the model. That is, the goal is to produce some model with similar statistics (as measured, for instance, by KL-divergence) on the observables as the model that generated the samples. This does not necessarily guarantee similarity in the parameter space of the models. In this paper, we focus on the stricter goal of *identifying* the model. That is, our goal is to produce a model whose parameters are sufficiently close to the true underlying model to generate also similar statistics on the observables. It is known that identification is not always possible, as there exist some distributions on the observables that can be generated by more than one model. However, a mild condition of separability, namely that the distribution on each observable is different among the sub-populations, guarantees identifiability information theoretically (Tahmasebi et al., 2018). We shall use ζ to denote the minimum difference between sub-population distributions on an observable. This will be defined precisely later. Note that identification also implies learning.

As pointed out in Feldman et al. (2008) and elsewhere, the identification problem in the case of observables taking values in a finite set reduces to the case of binary observables. Moreover, it is known from Fan and Li (2022) that with polynomial overhead in the size of the range of the observables, the problem reduces to the case of identifying the conditional expectations of real-valued observables that are independent conditional on the sub-population. Thus, we shall focus in this paper on this real-valued version of the problem.

Our main result is an algorithm that identifies the parameters of any ζ-separated mixture of product distributions using $(1/\zeta)^{O(k)}(1/\pi_{\min})^{O(1)}(1/\varepsilon)^2$ samples and runtime, up to additive error ε. Here $\pi_{\min}$ is the minimum frequency of a sub-population. The result holds if there are at least $2k-1$ ζ-separated observables, which is known to be a necessary condition (Teicher, 1961; Blischke, 1964). This result greatly improves upon Gordon et al. (2021), the best previously known complexity for identification (and learning), which required $(1/\zeta)^{O(k^2 \log k)}(1/\pi_{\min})^{O(\log k)}(1/\varepsilon)^2$ samples from $3k-3$ ζ-separated observables. Furthermore, we show that the sample complexity of identification (for constant ε) is at least $(1/(k\zeta))^{\Omega(k)}$ (note that $\zeta \leq \frac{1}{k-1}$ always). This generalizes the previously known lower bound of $\exp(-\Omega(k))$ that held only for $\zeta = \Theta(1/k)$ (Rabani et al., 2014). Hence, our results are essentially optimal, both in terms of the number of ζ-separated observables needed and in terms of sample and runtime complexity (excluding the case of $\zeta = \frac{1}{k^{1+o(1)}}$, where a small gap remains).

For large $n = \omega(k)$, if a subset of $2k-1$ ζ-separated observables is known, then the runtime bounds pick up an additional factor of n (to identify all the remaining observables). Otherwise, if the required subset exists but is not known, then the runtime picks up an additional factor of $n^{O(k)}$ (to enumerate over all possibilities). In both cases, the sample complexity only increases by a factor of $\log n$.

1.2. Related work and motivation.

The seminal work of Feldman et al. (2008) solves the learning problem for general k in sample and runtime complexity $n^{O(k^3)}$. This was improved in Chen and Moitra (2019) to $k^{O(k^3)}n^{O(k^2)}$.[1] Gordon et al. (2021) gave the first algorithm for the *identification problem*, identifying a k-component mixture on $3k-3$ ζ-separated variables with sample and runtime complexity of $(1/\zeta)^{O(k^2 \log k)}$. This was achieved by reducing the problem to a special case of identifying a mixture of k distributions on independent and *identically distributed* bits. The latter problem is solved using an elegant two-century-old method of Prony (de Prony, 1795) coupled with a robustness analysis given in Gordon et al. (2020).

The study of discrete mixture models is motivated by wide-ranging applications in fields such as population genetics, text classification and image recognition, see e.g. Pritchard et al. (2000); Juan and

1. Chen and Moitra (2019) also studied the problem of learning a "mixture of subcubes" of the hypercube, which is the special case where each random bit X_i is either fixed or uniformly distributed; for this case, they showed sample and runtime complexity of $n^{O(\log k)}$.

Vidal (2002, 2004); Li et al. (2016). Identification is often essential for these applications, particularly in the context of causal inference. Here, when data is drawn from multiple sources or sub-populations it is said to contain a *latent class* E. S. Allman (2009), which mirrors an unidentified mixture source. Standard procedures for identifying causal effects require considering the distributions within each latent class separately to control for potential confounding effects Pearl (2009). Such an approach is generally impossible unless these within-source probability distributions can be identified.

More broadly, the theory of causal inference relies at its core upon *Bayesian networks* of random variables (Pearl, 1985; Spirtes et al., 2000; Pearl, 2009); such a network imposes conditional independencies among random variables of the system. An important scenario is that several latent classes are subject to the same "system mechanics" (i.e., Bayesian network), but have different statistics. In this case, the problem of identifying the model is a far-reaching generalization of the problem of identifying mixtures of product distributions. Gordon et al. (2023) recently proposed an algorithm for this more general problem that uses the identification of mixtures of product distributions as an essential complexity-bottle-necking sub-routine. Thus the improvements of the present paper carry over directly to that application.

Note that in the context of Bayesian networks, the identification problem is already interesting for mixtures with just a few latent classes. This implies that the present algorithm can be useful in practice, despite the exponential dependence of its sample complexity on k (and furthermore, this dependence might not be exponential "in general", see the discussion in section 6).

1.3. Our methods.

We study the so-called Hadamard extensions that were also used to derive sample complexity in Gordon et al. (2021). We give a new and much more powerful bound on the condition number of the Hadamard extensions. This bound alone would improve the sample complexity of the algorithm in Gordon et al. (2021) to $(1/\zeta)^{O(k \log k)}$. We gain further improvement as follows: Instead of reducing the problem to identifying mixtures of iid (synthetic) bits, we reduce the problem to a tensor decomposition problem, where the tensor components are guaranteed to be well-conditioned (which makes the decomposition unique). An algorithm for tensor decomposition in this setting was given thirty years ago in Leurgans et al. (1993) and later analyzed for robustness in Goyal et al. (2014); Bhaskara et al. (2014). This algorithm can also be seen as a generalization of the matrix pencil method Hua and Sarkar (1990), applied to the iid case in Kim et al. (2019). Adapting the tensor decomposition algorithm to our setting and analyzing it using our new condition number bound, then yields the sample complexity of $(1/\zeta)^{O(k)}$.

1.4. Comparison with the parametric case.

The literature on mixture models for parametric families (exponential distributions, Gaussians in $\mathbb{R}$ or $\mathbb{R}^d$, etc.) is even more extensive and older than for discrete mixture models. It is essential to realize a fundamental difference between the types of problems. In general, data is generated by (unseen) selection of a sub–population j $(1 \leq j \leq k)$, followed by (seen) sampling of n independent samples *from the j-th distribution*. In almost every parametric scenario (think e.g., of a mixture of k Gaussians or exponential distributions on the line), $n = 1$ *is sufficient* in order to (in the limit of many repetitions) identify the model. This is fundamentally untrue in the non-parametric case; we have already mentioned that a lower bound of $n \geq 2k - 1$ was shown in Rabani et al. (2014); this threshold for n is called there the "aperture" of the problem. To see, for starters, why the aperture must be larger than $n = 1$, consider a single binary variable with $k = 2$ equiprobable sources (i.e., $\Pr(U = 0) = \Pr(U = 1) = 1/2$), one of which has $\Pr(X = 1 \mid U = 0) = \frac{3}{4}$ and the other $\Pr(X = 1 \mid U = 1) = \frac{1}{4}$. If we see after each selection of a source only a single sample of X, it is impossible to distinguish between the above mixture and a mixture in which $\Pr(X = 1 \mid U = 0) = 1$ and $\Pr(X = 1 \mid U = 1) = 0$. With access to multiple independent samples from

the *same* source, however, we get empirical estimates of higher moments of the distribution, and at the critical aperture can identify the model.

1.5. Organization.

Section 2 formally states the identification problem for mixtures of product distributions and sets up the mathematical objects needed for our work. Section 3 describes our algorithm and states our upper bounds on its sample complexity. These bounds mainly rely on our key result about the condition number of Hadamard extensions, which we prove in Section 4, with a full end-to-end analysis of our algorithm in Appendix A. Section 5 introduces the idea for our lower bounds on the complexity of the identification problem, with full proofs given in Appendix B. Finally, in section 6 we discuss potential further directions of research on our topic.

2. Results and preliminaries

2.1. The k-MixProd problem

Consider n real random variables $X_1, \dots, X_n$ that are supported on $[0,1]$ and are independent conditional on a latent random variable U with range $[k] = \{1, \dots, k\}$. Given iid samples of the joint distribution of $(X_1, \dots, X_n)$, we want to identify the distribution of U, given by $\pi_j := \Pr(U = j)$ $(j \in [k])$, and the conditional expectations $\mathbf{m}_{ij} = \mathbb{E}(X_i \mid U = j)$ $(i \in [n], j \in [k])$. Hence, the model parameters for our problem are given by a vector $(\pi, \mathbf{m}) \in \Delta^{k-1} \times [0,1]^{n \times k}$, where Δ^{k-1} denotes the $(k-1)$-simplex.

Set $X_S = \prod_{i \in S} X_i$, so $\mathbb{E}(X_S \mid U = j) = \prod_{i \in S} \mathbf{m}_{ij}$. The mapping of the model to the statistics is then given by:

$$\gamma_n : \Delta^{k-1} \times [0,1]^{n \times k} \to \mathbb{R}^{2^{[n]}}$$
$$\gamma_n(\pi, \mathbf{m})(S) = \mathbb{E}(X_S) = \sum_{j=1}^{k} \pi_j \, \mathbb{E}(X_S \mid U = j) = \sum_{j=1}^{k} \pi_j \prod_{i \in S} \mathbf{m}_{ij}$$

We drop the subscript n and write γ when n is implied. The k-MixProd identification problem is to invert γ_n, i.e., to recover $(\pi_j)_{j \in [k]}$ and $(\mathbf{m}_{ij})_{i \in [n], j \in [k]}$ (up to permuting the set $[k]$). This task is interesting in two versions, exact identification of $(\pi, \mathbf{m})$ from $\gamma_n(\pi, \mathbf{m})$ (i.e., from perfect statistics), and approximate identification of $(\pi, \mathbf{m})$ from noise-perturbed statistics $\tilde{\mathrm{g}}$, i.e., from $\tilde{\mathrm{g}} \in \mathbb{R}^{\{0,1\}^n}$ that is close to $\gamma_n(\pi, \mathbf{m})$. To make the latter goal precise we need to specify metrics on the domain and range of γ_n. These are L_∞ metrics, up to relabelings of the latent variable. (S_k denotes the symmetric group on k letters.)

$$d_{\mathrm{model}}((\pi, \mathbf{m}), (\pi', \mathbf{m}')) := \min_{\rho \in S_k} \max\{\max_j |\pi_j - \pi'_{\rho(j)}|, \max_{i,j} |\mathbf{m}_{i,j} - \mathbf{m}'_{i,\rho(j)}|\}$$
$$d_{\mathrm{stat}}(\mathrm{g}, \mathrm{g}') := \max_{S \subseteq [n]} |\,\mathrm{g}(S) - \mathrm{g}'(S)|.$$

The mapping γ_n is not everywhere injective, so the k-MixProd model identification problem is not always feasible. To guarantee identifiability we need to make the following assumptions:

(a) (ζ-separation) each variable X_i is ζ-*separated*, i.e. $|\mathbf{m}_{ij} - \mathbf{m}_{ij'}| \geq \zeta$ for all $j \neq j' \in [k]$;

(b) (non-degenerate prior) for each $j \in [k]$, we have $\pi_j \geq \pi_{\min} > 0$;

(c) (sufficiently many observables) there are at least $n \geq 2k - 1$ variables X_i.

Let $\mathcal{D}_{n,\zeta,\pi_{\min}}$ denote the space of the k-MixProd models with n variables satisfying assumptions (a) and (b). Formally:

$$\mathcal{D}_{n,\zeta,\pi_{\min}} = \{(\pi, \mathbf{m}) \in \Delta^{k-1} \times [0,1]^{n\times k} \mid \min_j \pi_j \geq \pi_{\min},\ \forall i\ \min_{j\neq j'} |\mathbf{m}_{ij} - \mathbf{m}_{ij'}| \geq \zeta\},$$

Theorem 10 shows (quite apart from its algorithmic content) that for $n \geq 2k-1$, if $(\pi, \mathbf{m})$ is a model¡ in $\mathcal{D}_{n,\zeta,\pi_{\min}}$, then any model whose statistics are close (in d_{stat}) to those of $(\pi, \mathbf{m})$, must also be close to $(\pi, \mathbf{m})$ in d_{model}.

We now introduce some mathematical concepts that will be needed for our work.

2.2. Hadamard extensions and multilinear moments

First, we define some of our notation for working with matrices throughout the paper.

Definition 1 *Given a matrix $\mathbf{A}$ of any dimensions, let $\mathbf{A}_{i*}$ denote the i'th row of $\mathbf{A}$, and $\mathbf{A}_{*j}$ the j'th column of $\mathbf{A}$. Where clear from context we write $\mathbf{A}_i$ instead of $\mathbf{A}_{i*}$. For S a set of rows, $\mathbf{A}[S]$ denotes the submatrix of $\mathbf{A}$ consisting of the rows in S.*

Definition 2 *The singular values of a real matrix $\mathbf{A}$ are denoted $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots$. The $L_2 \to L_2$ operator norm is denoted $\|\mathbf{A}\|$. The condition number of $\mathbf{A}$ is denoted $\kappa(\mathbf{A}) = \|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\|$.*

The key mathematical objects we are studying are given by the following definitions:

Definition 3 (Hadamard product) *The* Hadamard product *is the mapping $\odot : \mathbb{R}^{[k]} \times \mathbb{R}^{[k]} \to \mathbb{R}^{[k]}$ which, for row vectors $u = (u_1, \ldots, u_k)$ and $v = (v_1, \ldots, v_k)$, is given by $u \odot v \coloneqq (u_1 v_1, \ldots, u_k v_k)$. Equivalently, using the notation $v_\odot = \operatorname{diag}(v)$, the Hadamard product is $u \odot v = u \cdot v_\odot$. The identity element for the Hadamard product is the all-ones row vector $\mathbb{1}$.*

Definition 4 (Hadamard extension) *For $\mathbf{n} \in \mathbb{R}^{n\times p}$, the* Hadamard extension *of $\mathbf{n}$, written $\mathbb{H}(\mathbf{n})$, is the $2^n \times p$ matrix with rows $\mathbb{H}(\mathbf{n})_S$ for all $S \subseteq [n]$, where, for $S = \{i_1, \ldots, i_\ell\}$, $\mathbb{H}(\mathbf{n})_S = \mathbf{n}_{i_1} \odot \cdots \odot \mathbf{n}_{i_\ell}$; equivalently $\mathbb{H}(\mathbf{n})_{S,j} = \prod_{i\in S} \mathbf{n}_{ij}$. In particular $\mathbb{H}(\mathbf{n})_\emptyset = \mathbb{1}$, and for all $i \in [n]$, $\mathbb{H}(\mathbf{n})_{\{i\}} = \mathbf{n}_i$.*

To our knowledge, this construction first appeared (not under this name) in Chen and Moitra (2019). Recall that the data we obtain from our samples will be estimates of $\mathbb{E}[X_S] = \mathbb{E}[\prod_{i\in S} X_i]$ for all subsets $S \subseteq [n]$. We call these the *multilinear moments* of the distribution, since they are multilinear in the rows $\mathbf{m}_i$. Observe that $\mathbb{E}[X_S] = \sum_j \pi_j \prod_{i\in S} \mathbf{m}_{ij} = (\mathbb{H}(\mathbf{m}))_S \cdot \pi$, or equivalently, $\mathbb{E}[X_S] = (\mathbf{m}_{i_1} \odot \mathbf{m}_{i_2} \odot \cdots \odot \mathbf{m}_{i_s})\pi$ where $S = \{i_1, i_2, \ldots, i_s\}$. Hence, the vector of statistics for a model $(\pi, \mathbf{m})$ is given by $\gamma(\pi, \mathbf{m}) = \mathbb{H}(\mathbf{m})\pi$, motivating the study of Hadamard extensions for this problem. We can see immediately that source identification is not possible if $\mathbb{H}(\mathbf{m})$ has less than full column rank, i.e., if $\operatorname{rank}(\mathbb{H}(\mathbf{m})) < k$, as then the mixing weights cannot be unique. We will organize our observed statistics into the following objects, which we will use for identification:

Definition 5 *Given disjoint sets $S, T \subseteq \{2, \ldots, n\}$, define*

$$\mathbf{C}_{ST} = \mathbb{H}(\mathbf{m}[S]) \cdot \pi_\odot \cdot \mathbb{H}(\mathbf{m}[T])^\mathsf{T},$$
$$\mathbf{C}_{ST,1} = \mathbb{H}(\mathbf{m}[S]) \cdot \pi_\odot \cdot \mathbf{m}_{1\odot} \cdot \mathbb{H}(\mathbf{m}[T])^\mathsf{T}$$

Note, for $A \subseteq S, B \subseteq T$,

$$(\mathbf{C}_{ST})_{A,B} = \gamma_n(\pi, \mathbf{m})(A \cup B),$$
$$(\mathbf{C}_{ST,1})_{A,B} = \gamma_n(\pi, \mathbf{m})(A \cup B \cup \{1\})$$

Consequently $\mathbf{C}_{ST}$ and $\mathbf{C}_{ST,1}$ are observable, that is, every one of their entries is a statistic which the algorithm receives (a noisy version of) as input.

2.3. Tensor decomposition

A matrix $\mathbf{A}$ is rank 1 if and only if it can be written as $\mathbf{A} = \mathbf{u}\mathbf{v}^\mathsf{T}$ for some vectors $\mathbf{u}, \mathbf{v}$. This concept can be generalized for tensors:

Definition 6 *A 3-way tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is said to be of rank 1 if there exist vectors $u \in \mathbb{R}^{d_1}, v \in \mathbb{R}^{d_2}, z \in \mathbb{R}^{d_3}$ such that for all i, j, k:*

$$\mathcal{T}_{ijk} = u_i \cdot v_j \cdot z_k.$$

Now, the rank of any tensor $\mathcal{T}$ can be defined as the minimum number of rank-1-tensors that sum up to $\mathcal{T}$. Equivalently, a tensor of rank r has the following decomposition:

Definition 7 *A 3-way tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ has a rank-r-decomposition if there exist matrices $\mathbf{U} \in \mathbb{R}^{d_1 \times r}, \mathbf{V} \in \mathbb{R}^{d_2 \times r}, \mathbf{Z} \in \mathbb{R}^{d_3 \times r}$ such that for all i, j, k:*

$$\mathcal{T}_{ijk} = \sum_{\ell=1}^{r} \mathbf{U}_{i\ell}\mathbf{V}_{j\ell}\mathbf{Z}_{k\ell}.$$

We write $\mathcal{T} = [\mathbf{U}, \mathbf{V}, \mathbf{Z}]$ and call $\mathbf{U}, \mathbf{V}, \mathbf{Z}$ the factor matrices or tensor components of $\mathcal{T}$.

In general, the rank-r-decomposition of a tensor $\mathcal{T}$ need not be unique, but a classical result in Kruskal (1977) gives sufficient conditions for uniqueness.

Definition 8 *The Kruskal rank of a matrix $\mathbf{A}$ is the largest number r such that any r columns of $\mathbf{A}$ are linearly independent.*

Theorem 9 (Kruskal (1977)) *The rank-r-decomposition of a three-way tensor $\mathcal{T} = [\mathbf{U}, \mathbf{V}, \mathbf{Z}]$ is unique up to scaling and permuting the columns of the factor matrices if*

$$k_{\mathbf{U}} + k_{\mathbf{V}} + k_{\mathbf{Z}} \geq 2r + 2,$$

where $k_{\mathbf{U}}, k_{\mathbf{V}}, k_{\mathbf{Z}}$ denote the Kruskal rank of the matrices $\mathbf{U}, \mathbf{V}, \mathbf{Z}$ respectively.

3. The algorithm

3.1. Reducing k-MixProd to tensor decomposition

To motivate our algorithm, we first discuss a way of solving the k-MixProd identification problem for $n \geq 2k - 1$ given perfect statistics. Consider three disjoint sets $S, T, U \subseteq [n]$ of ζ-separated observables, such that $|S| = |T| = k - 1$ and $|U| = 1$. For convenience, we index rows so that $U = \{1\}$. Consider the vector of perfect statistics $\mathrm{g} = \gamma_{2k-1}(\mathbf{m}[S \cup T \cup \{1\}])$ that corresponds to the observables in $S \cup T \cup \{1\}$. We can naturally view g as a three-way tensor $\mathcal{T} \in \mathbb{R}^{2^S \times 2^T \times 2^{\{1\}}}$ whose entries are given by $\mathcal{T}_{A,B,C} = \mathrm{g}_{A \cup B \cup C}$ for $A \subseteq S, B \subseteq T, C \subseteq \{1\}$. Since we have

$$\mathcal{T}_{A,B,C} = \sum_{j=1}^{k} \pi_j \prod_{i \in A \cup B \cup C} \mathbf{m}_{ij} = \sum_{j=1}^{k} \mathbb{H}(\mathbf{m}[S])_{A,j} \cdot \mathbb{H}(\mathbf{m}[T])_{B,j} \cdot (\mathbb{H}(\mathbf{m}_1) \cdot \pi_\odot)_{C,j},$$

the tensor $\mathcal{T}$ can be decomposed as $\mathcal{T} = [\mathbb{H}(\mathbf{m}[S]), \mathbb{H}(\mathbf{m}[T]), \mathbb{H}(\mathbf{m}_1)\pi_\odot]$. It will follow as a "qualitative" corollary of Theorem 13 that $\mathbb{H}(\mathbf{m}[S])$ and $\mathbb{H}(\mathbf{m}[T])$ have full column rank, which implies both matrices have Kruskal rank k. Moreover, ζ-separation of X_1 implies that the matrix $\mathbb{H}(\mathbf{m}_1) \cdot \pi_\odot$ has Kruskal

rank 2. Hence, by Theorem 9, the decomposition of $\mathcal{T}$ is unique. We deduce that identifying the model parameters $\mathbf{m}$ and π from perfect statistics (provided in $\mathcal{T}$) is equivalent to computing the unique tensor decomposition of $\mathcal{T}$. An efficient algorithm for computing tensor decomposition in this setting has first been given by Leurgans et al. (1993) and later analyzed for stability by Goyal et al. (2014) and Bhaskara et al. (2014). The idea is to first project the components of $\mathcal{T}$ down to their image, i.e. find matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{2^{k-1} \times k}$ such that $\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])$ and $\mathbf{V}^\mathsf{T}\mathbb{H}(\mathbf{m}[T])$ are invertible. Then, compute $\hat{\mathbf{C}}_{ST} = \mathbf{U}^\mathsf{T}\mathbf{C}_{ST}\mathbf{V}$ and $\hat{\mathbf{C}}_{ST,1} = \mathbf{U}^\mathsf{T}\mathbf{C}_{ST,1}\mathbf{V}$ (see Definition 5) from the given statistics. Now, the key observation is that the tensor components $\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])$ and $\mathbf{V}^\mathsf{T}\mathbb{H}(\mathbf{m}[T])$ can be found as the eigenvectors of $\hat{\mathbf{C}}_{ST,1}\hat{\mathbf{C}}_{ST}^{-1}$ and $\hat{\mathbf{C}}_{ST,1}^\mathsf{T}(\hat{\mathbf{C}}_{ST}^\mathsf{T})^{-1}$ respectively. This is because

$$\begin{aligned}
\hat{\mathbf{C}}_{ST,1}\hat{\mathbf{C}}_{ST}^{-1} &= \mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]) \cdot \pi_\odot \cdot \mathbf{m}_{1\odot} \cdot \mathbb{H}(\mathbf{m}[T])^\mathsf{T}\mathbf{V} \cdot \left(\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]) \cdot \pi_\odot \cdot \mathbb{H}(\mathbf{m}[T])^\mathsf{T}\mathbf{V}\right)^{-1} \\
&= \mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]) \cdot \pi_\odot \cdot \mathbf{m}_{1\odot} \cdot \mathbb{H}(\mathbf{m}[T])^\mathsf{T}\mathbf{V} \cdot \left(\mathbb{H}(\mathbf{m}[T])^\mathsf{T}\mathbf{V}\right)^{-1} \cdot \pi_\odot^{-1} \cdot \left(\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])\right)^{-1} \\
&= \mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]) \cdot \mathbf{m}_{1\odot} \cdot \left(\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])\right)^{-1}, \qquad (1)
\end{aligned}$$

and

$$\begin{aligned}
\hat{\mathbf{C}}_{ST,1}^\mathsf{T}(\hat{\mathbf{C}}_{ST}^\mathsf{T})^{-1} &= \left(\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]) \cdot \pi_\odot \cdot \mathbf{m}_{1\odot} \cdot \mathbb{H}(\mathbf{m}[T])^\mathsf{T}\mathbf{V}\right)^\mathsf{T} \cdot \left(\left(\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]) \cdot \pi_\odot \cdot \mathbb{H}(\mathbf{m}[T])^\mathsf{T}\mathbf{V}\right)^\mathsf{T}\right)^{-1} \\
&= \mathbf{V}^\mathsf{T}\mathbb{H}(\mathbf{m}[T]) \cdot \pi_\odot \cdot \mathbf{m}_{1\odot} \cdot \mathbb{H}(\mathbf{m}[S])^\mathsf{T}\mathbf{U} \cdot \left(\mathbb{H}(\mathbf{m}[S])^\mathsf{T}\mathbf{U}\right)^{-1} \cdot \pi_\odot^{-1} \cdot \left(\mathbf{V}^\mathsf{T}\mathbb{H}(\mathbf{m}[T])\right)^{-1} \\
&= \mathbf{V}^\mathsf{T}\mathbb{H}(\mathbf{m}[T]) \cdot \mathbf{m}_{1\odot} \cdot \left(\mathbf{V}^\mathsf{T}\mathbb{H}(\mathbf{m}[T])\right)^{-1}. \qquad (2)
\end{aligned}$$

In both cases, the eigenvalues of the matrices above are given by the entries of $\mathbf{m}_1$. Crucially, ζ-separation of $\mathbf{m}_1$ allows us to match up the columns of $\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])$ and $\mathbf{V}^\mathsf{T}\mathbb{H}(\mathbf{m}[T])$ (and guarantees numerical stability). The original entries of $\mathbb{H}(\mathbf{m})$ and π can then be found from linear systems. In fact, notice that

$$\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])\pi = \mathbf{U}^\mathsf{T}\,\mathrm{g}[2^S], \qquad (3)$$

and for any row $\mathbf{m}_i$ with $i \notin S$, we have

$$\mathbf{U}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])\pi_\odot\mathbf{m}_i^\mathsf{T} = \mathbf{U}^\mathsf{T}\,\mathrm{g}(R \cup \{i\})_{R\subseteq S}. \qquad (4)$$

At this point, π and $\mathbf{m}_i$ are the only unknowns in the equations above, so we have solved the identification problem. Our main result essentially states that this identification method is robust to noise, i.e. it still works when only presented with approximate statistics. We formally state the identification algorithm as Algorithm 1.

Some comments on the algorithm: first, it is not necessary that the sets S, T are of size $k-1$; typically, $\lceil \lg k \rceil$ will suffice. It is not even necessary that observables are ζ-separated. These assumptions guarantee that $\sigma_k(\mathbf{C}_{ST})$ is large, but the latter, along with the argument that this is w.h.p. reproduced for $\sigma_k(\tilde{\mathbf{C}}_{ST})$, is sufficient for the success of the algorithm.

Second, we do not need to start with knowledge of S, T. Given n variables of which an unknown subset of $2k-1$ variables are ζ-separated, we can simply perform the algorithm for all possible choices of subsets S, T of size $k-1$ (and an additional single row). Then, we choose the computed model whose statistics are closest to the observed statistics as the final output. This exhaustive search can increase the runtime by a factor of about n^{2k}; but actually all these complexities are only exponential in the *actual* needed size of $S \cup T$. As noted, for generic $\mathbf{m}$ will be as small as $\lceil \lg k \rceil$, which makes the algorithm far more attractive in practice.

Algorithm 1 (adapted from Leurgans et al. (1993), Bhaskara et al. (2014)) Identifies a mixture of product distributions on $2k-1$ binary variables given the joint distribution.

Input: Two disjoint sets S, T of $\zeta-$separated, binary observables X_i, each of cardinality $k-1$; a single $\zeta-$separated, binary observable that is not part of S, T, wlog X_1; vector $\tilde{\mathrm{g}} \in \mathbb{R}^{2^{S \cup T \cup \{1\}}}$, with $\tilde{\mathrm{g}}(R)$ the empirical approximation to $E(X_R) = \gamma(\pi, \mathbf{m})(R)$).

Construct $\tilde{\mathbf{C}}_{ST}$ and $\tilde{\mathbf{C}}_{ST,1}$ by $(\tilde{\mathbf{C}}_{ST})_{AB} = \tilde{\mathrm{g}}(A \cup B)$ and $(\tilde{\mathbf{C}}_{ST,1})_{AB} = \tilde{\mathrm{g}}(A \cup B \cup \{1\})$ **for** $A \subseteq S, B \subseteq T$.

Set $\hat{\mathbf{U}} \in \mathbb{R}^{2^{|T|} \times k}$ to be the top k left singular vectors of $\tilde{\mathbf{C}}_{ST}$ and $\hat{\mathbf{V}} \in \mathbb{R}^{2^{|S|} \times k}$ to be the top k right singular vectors of $\tilde{\mathbf{C}}_{ST}$.

$\hat{\mathbf{C}}_{ST} \leftarrow \hat{\mathbf{U}}^{\mathsf{T}} \tilde{\mathbf{C}}_{ST} \hat{\mathbf{V}}$, $\hat{\mathbf{C}}_{ST,1} \leftarrow \hat{\mathbf{U}}^{\mathsf{T}} \tilde{\mathbf{C}}_{ST,1} \hat{\mathbf{V}}$

Set $\hat{\mathbf{S}}$ to be the eigenvectors of $\hat{\mathbf{C}}_{ST,1}(\hat{\mathbf{C}}_{ST})^{-1}$ (sorted from highest eigenvalue to lowest).

Set $\hat{\mathbf{T}}$ to be the eigenvectors of $\hat{\mathbf{C}}_{ST,1}^{\mathsf{T}}(\hat{\mathbf{C}}_{ST}^{\mathsf{T}})^{-1}$ (sorted from highest eigenvalue to lowest).

$\tilde{\pi} \leftarrow \hat{\mathbf{S}}^{-1} \cdot \hat{\mathbf{U}}^{\mathsf{T}} \left(\tilde{\mathrm{g}}(R)_{R \subseteq S}\right)$

for every $i \in T \cup \{1\}$, $\tilde{\mathbf{m}}_i \leftarrow \left((\tilde{\mathrm{g}}(R \cup \{i\}))_{R \subseteq S}\right)^{\mathsf{T}} \cdot \hat{\mathbf{U}} \cdot \left(\hat{\mathbf{S}}^{\mathsf{T}}\right)^{-1} \cdot \tilde{\pi}_{\odot}^{-1}$.

for every $i \in S$, $\tilde{\mathbf{m}}_i \leftarrow \left((\tilde{\mathrm{g}}(R \cup \{i\}))_{R \subseteq T}\right)^{\mathsf{T}} \cdot \hat{\mathbf{V}} \cdot \left(\hat{\mathbf{T}}^{\mathsf{T}}\right)^{-1} \cdot \tilde{\pi}_{\odot}^{-1}$.

Theorem 10 *Let $n = 2k-1$ and fix any $\varepsilon \in (0, \zeta/2)$. Let $(\pi, \mathbf{m}) \in \mathcal{D}_{n,\zeta,\pi_{\min}}$. If Algorithm 1 is given approximate statistics $\tilde{\mathrm{g}}$ on $(X_1, \ldots, X_n)$ as input, satisfying $d_{\mathrm{stat}}(\gamma(\pi, \mathbf{m}), \tilde{\mathrm{g}}) < {\pi_{\min}}^{O(1)} \zeta^{O(k)} \varepsilon$, then in runtime $\exp(O(k))$ the algorithm outputs $(\tilde{\pi}, \tilde{\mathbf{m}})$ s.t.*

$$d_{\mathrm{model}}((\pi, \mathbf{m}), (\tilde{\pi}, \tilde{\mathbf{m}})) < \varepsilon. \tag{5}$$

Moreover, this output is essentially unique, in the sense that: any (not necessarily ζ-separated) model $(\pi', \mathbf{m}')$ with $d_{\mathrm{stat}}(\gamma(\pi, \mathbf{m}), \gamma(\pi', \mathbf{m}')) < {\pi_{\min}}^{O(1)} \zeta^{O(k)} \varepsilon$ also satisfies (5).

Corollary 11 *For $n = 2k-1$ random variables X_i and $\delta \in (0,1)$, sample complexity*

$$(1/\zeta)^{O(k)} (1/\pi_{\min})^{O(1)} (1/\varepsilon)^2 \log(1/\delta)$$

suffices to compute a model that satisfies (5) with probability at least $1-\delta$.

Corollary 12 *Let the number of observables be $n \geq 2k-1$, and of these let some $2k-1$ be ζ-separated. If this subset is known, then sample complexity*

$$\log n \cdot (1/\zeta)^{O(k)} (1/\pi_{\min})^{O(1)} (1/\varepsilon)^2 \log(1/\delta)$$

and post-sampling runtime $n \cdot \exp(O(k))$ suffices to compute a model that satisfies (5) with probability at least $1-\delta$. If the subset is not known, then the same sample complexity and post-sampling runtime $n^{2k} \cdot \exp(O(k))$ suffices to compute a model that w.h.p. satisfies (5).

Corollary 11 follows from Theorem 10 by standard Chernoff bounds and a union bound. In the following, we give an overview of the steps necessary to analyze Algorithm 1 and therefore prove Theorem 10. Full proofs of Theorem 10 and Corollary 12 can be found in Appendix A.

3.2. Analyzing robustness

This section describes what each non-trivial line of Algorithm 1 accomplishes, and in each case points to the part of the analysis necessary to justify it. We first set some definitions for the analysis. Let $n = 2k - 1$, let $\{2, \ldots, n\}$ be the disjoint union of sets S, T each of size $k - 1$. Let $\mathrm{g} = \gamma_n(\pi, \mathbf{m})$ and let $\tilde{\mathrm{g}} \in \mathbb{R}^{\{0,1\}^n}$ be the empirical statistics. The analysis relies on assuming that $(\pi, \mathbf{m}) \in \mathcal{D}_{n,\zeta,\pi_{\min}}$. In what follows let

$$\mathbf{d} := d_{\mathrm{stat}}(\tilde{\mathrm{g}}, \mathrm{g}). \tag{6}$$

Now for the line-by-line:

- Line 3: The SVD of $\tilde{\mathbf{C}}_{ST}$ can be computed with high numerical accuracy in time $\exp(O(k))$ using, for instance, the Golub-Kahan-Reinsch algorithm (see Lemma 20 in the appendix and Golub and Loan (2013), Chapter 8). The span of the top k left and right singular vectors of $\tilde{\mathbf{C}}_{ST}$ approximate the images of $\mathbb{H}(\mathbf{m}[S])$ and $\mathbb{H}(\mathbf{m}[T])$ respectively.
- Line 4: We project on the top k singular vectors of $\tilde{\mathbf{C}}_{ST}$ to get an invertible matrix $\hat{\mathbf{C}}_{ST}$. Lemma 21 bounds the condition number of $\hat{\mathbf{C}}_{ST}$, given that $\mathbf{d}$ is small enough. Note that the two matrices $\hat{\mathbf{C}}_{ST}$ and $\hat{\mathbf{C}}_{ST,1}$ can be arranged to a $k \times k \times 2$-tensor $\hat{\mathcal{T}}$ that is close to the tensor $\mathcal{T} = \left[\hat{\mathbf{U}}^{\mathsf{T}}\mathbb{H}(\mathbf{m}[S]), \hat{\mathbf{V}}^{\mathsf{T}}\mathbb{H}(\mathbf{m}[T]), \mathbb{H}(\mathbf{m}_1)\pi_{\odot}\right]$ by Lemma 22.
- Lines 5,6: These two lines implement the core of the tensor decomposition algorithm from Leurgans et al. (1993), as explained in the previous section. Hence, the matrices $\hat{\mathbf{S}}$ and $\hat{\mathbf{T}}$ approximate the tensor components $\hat{\mathbf{U}}^{\mathsf{T}}\mathbb{H}(\mathbf{m}[S])$ and $\hat{\mathbf{V}}^{\mathsf{T}}\mathbb{H}(\mathbf{m}[T])$ of $\mathcal{T}$. The diagonalization steps can be performed with high numerical accuracy in time $\exp(O(k))$ using, for instance, the algorithm from Banks et al. (2020). For the error bounds, we rely on Theorem 23 from Bhaskara et al. (2014).
- Lines 7-9: Here, we solve for the model parameters π and $\mathbf{m}$, using the equations (3) and (4). For error control, we apply Lemma 24.

Essentially, all the analysis steps above go through given that the distance $\mathbf{d}$ between the empirical and the true statistics is bounded by $\mathrm{poly}(2^{-k}, \sigma_k(C_{ST}))^{-1})$ (the first argument comes from the inverse of the dimension of C_{ST} and will be dominated by the second argument). Hence, the sample complexity is basically a polynomial function in the k'th singular value of C_{ST}, which we will analyze in the next section.

4. The condition number bound

The key to the sample-complexity and runtime bounds for our algorithm lies in the following condition number bound for the Hadamard Extension.

Theorem 13

1. *Let $\mathbf{m}$ consist of $k-1$ ζ-separated rows in $\mathbb{R}^k$, and observe that the singular values of $\mathbb{H}(\mathbf{m})$ satisfy $\sigma_1(\mathbb{H}(\mathbf{m})) \geq \ldots \geq \sigma_k(\mathbb{H}(\mathbf{m})) \geq 0 = \sigma_{k+1}(\mathbb{H}(\mathbf{m})) = \ldots = \sigma_{2^{k-1}}(\mathbb{H}(\mathbf{m}))$. Then*

$$\sigma_k(\mathbb{H}(\mathbf{m})) > \frac{1}{\sqrt{k}}\left(\frac{\zeta}{2\sqrt{5}}\right)^{k-1} =: \boldsymbol{\sigma}. \tag{7}$$

2. *Let $(\pi, \mathbf{m}) \in \mathcal{D}_{2k-1,\zeta,\pi_{\min}}$ and let $\mathbf{C}_{ST}$ be as in Defn. 5. Then*

$$\sigma_k(\mathbf{C}_{ST}) > \pi_{\min}\boldsymbol{\sigma}^2 = \frac{\pi_{\min}}{k}\left(\frac{\zeta}{2\sqrt{5}}\right)^{2k-2}.$$

Definition 14 $\mathbb{H}_p = \{\mathbb{H}(\mathbf{n}) : \mathbf{n} \in \mathbb{R}^{[k-1]\times[p]}\}$. *(So $\mathbb{H}_1$ consists of rank-1 tensors of order $k-1$.)*

The proof of Theorem 13 relies on the following insight. Since $\mathbb{H}(\mathbf{m})$ has dimensions $2^{k-1} \times k$, $\sigma_k(\mathbb{H}(\mathbf{m}))$ characterizes the least norm of $\mathbb{H}(\mathbf{m}) \cdot v$ ranging over any unit vector v (all norms in L_2), but it does *not* characterize the least norm of vectors of the form $h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m})$, which is 0 as the left-kernel of $\mathbb{H}(\mathbf{m})$ is of course very large. The insight is that it *does* become possible to bound $\sigma_k(\mathbb{H}(\mathbf{m}))$ in terms of such vectors h, *provided h is restricted to rank 1 tensors.* With this in mind we define:

$$\tau(\mathbf{m}) = \min_{0\neq h\in\mathbb{H}_1} \left\| h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m}) \right\| / \|h\|.$$

Proof of Theorem 13 To show Part 2 from Part 1: The SVD implies there is a k-dimensional space V s.t. $\forall v \in V$, $\|v^\mathsf{T} \cdot \mathbb{H}(\mathbf{m}[S])\| \geq \boldsymbol{\sigma}\|v\|$. Further, for all $w \in \mathbb{R}^k$, $\|w^\mathsf{T} \cdot \pi_\odot\| \geq \pi_{\min}\|w\|$. And for all $w \in \mathbb{R}^k$, $\|w^\mathsf{T} \cdot \mathbb{H}(\mathbf{m}[T])^\mathsf{T}\| \geq \boldsymbol{\sigma}\|w\|$. So $\forall v \in V$, $\|v \cdot \mathbf{C}_{TS}\| \geq \pi_{\min}\boldsymbol{\sigma}^2\|v\|$.

In order to establish Part 1 we prove the following two lemmas.

Lemma 15 $\sigma_k(\mathbb{H}(\mathbf{m})) \geq \tau(\mathbf{m})/\sqrt{k}$.

Lemma 16 $\tau(\mathbf{m}) > (\zeta/2\sqrt{5})^{k-1}$.

Proof of Lemma 15 Consider $v \in \mathbb{R}^k$, $\|v\| = 1$, achieving $\sigma_k(\mathbb{H}(\mathbf{m}))$, i.e., $r \coloneqq \mathbb{H}(\mathbf{m}) \cdot v$ satisfies $\|r\| = \sigma_k(\mathbb{H}(\mathbf{m}))$. W.l.o.g. the order of coordinates is such that $|v_k| \geq 1/\sqrt{k}$. The last column of $\mathbb{H}(\mathbf{m})$ is then:

$$\mathbb{H}(\mathbf{m})_{*k} = \frac{1}{v_k}\left(r - \sum_{j=1}^{k-1} v_j \mathbb{H}(\mathbf{m})_{*j}\right). \tag{8}$$

Now we carefully choose $h \in \mathbb{H}_1$ based on $\mathbf{m}$. Define the column vector $\mathbf{n} \in \mathbb{R}^{[k-1]}$ by $\mathbf{n}_i \coloneqq -1/\mathbf{m}_{ii}$ $(1 \leq i \leq k-1)$; and let

$$h \coloneqq \left(\prod_1^{k-1} \mathbf{m}_{ii}\right) \mathbb{H}(\mathbf{n}). \tag{9}$$

For $j \neq k$ we have:

$$h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m})_{*j} \quad = \quad \sum_S (-1)^{|S|}\left(\prod_{i\notin S} \mathbf{m}_{ii}\right)\left(\prod_{i\in S} \mathbf{m}_{ij}\right) = \prod_{i=1}^{k-1}(\mathbf{m}_{ii} - \mathbf{m}_{ij}) = 0.$$

So, $h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m})_{*j} = 0$ for $j = 1, \ldots, k-1$. For $j = k$, we apply (8) to evaluate $h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m})_{*k}$:

$$\left(h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m})\right)_k = \frac{1}{v_k} h^\mathsf{T} \cdot \left(r - \sum_{j=1}^{k-1} v_j \mathbb{H}(\mathbf{m})_{*j}\right) = \frac{1}{v_k}\left(h^\mathsf{T} \cdot r - \sum_{j=1}^{k-1} v_j h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m})_{*j}\right) = \frac{1}{v_k} h^\mathsf{T} \cdot r.$$

The norm of $h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m})$ is then upper bounded by

$$\begin{aligned}
\left\| h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m}) \right\| &= \left| (h^\mathsf{T} \cdot \mathbb{H}(\mathbf{m}))_k \right| \\
&= \frac{1}{v_k} h^\mathsf{T} \cdot r \\
&\leq \sqrt{k}\, \|h\|\, \|r\| \\
&= \sqrt{k}\, \|h\|\, \sigma_k(\mathbb{H}(\mathbf{m})).
\end{aligned}$$

■

Proof of Lemma 16 Consider any $\mathbf{G} \in \mathbb{H}_1$, say $\mathbf{G} = \mathbb{H}(\mathbf{g})$, $\mathbf{g} \in \mathbb{R}^{[k-1]}$. Then $(\mathbf{G}^\mathsf{T} \cdot \mathbb{H}(\mathbf{m}))_j = \sum_S \mathbf{G}_S \mathbb{H}(\mathbf{m})_{S,j} = \prod_1^{k-1}(1 + \mathbf{g}_i \mathbf{m}_{i,j})$. We also note that $\|\mathbf{G}\| = \sqrt{\prod_1^{k-1}(1+\mathbf{g}_i^2)}$.

We now show that there is some j such that $\prod_{i=1}^{k-1} \left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right|$ is large. First, for any i for which $\mathbf{g}_i \geq \frac{1}{2}$, there is at most one j s.t. $\mathbf{m}_{ij} \leq \zeta$; exclude these j's. Next, for each i for which $\mathbf{g}_i < \frac{1}{2}$, there is at most one j s.t. $\left|\frac{1}{\mathbf{g}_i} + \mathbf{m}_{ij}\right| \leq \zeta/2$; exclude these j's. For the remainder of the argument fix any j which has not been excluded. Since $\mathbf{m}$ has k columns while $\mathbf{g} \in \mathbb{R}^{k-1}$, such a j exists. We now lower bound $\left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right|$ for each i; there are three cases.

1. $\mathbf{g}_i \geq 1/2, \mathbf{m}_{ij} > \zeta$. Then $\left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right| \geq \mathbf{m}_{ij} > \zeta$.
2. $-1/2 < \mathbf{g}_i < 1/2$. Then $\left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right| > \sqrt{\frac{(\mathbf{m}_{ij}-2)^2}{5}} \geq 1/\sqrt{5}$.
3. $\mathbf{g}_i \leq -1/2, |\frac{1}{\mathbf{g}_i} + \mathbf{m}_{ij}| > \zeta/2$. Then $\left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right| = \left|\frac{\mathbf{g}_i(\frac{1}{\mathbf{g}_i}+\mathbf{m}_{ij})}{\mathbf{g}_i\sqrt{1+1/\mathbf{g}_i^2}}\right| > \frac{\zeta}{2\sqrt{5}}$.

We therefore have $\tau(\mathbf{m}) > (\zeta/2\sqrt{5})^{k-1}$. ■

Part 1 is an immediate consequence of the two lemmas. ■

5. Lower bounds

The following theorem shows that our algorithmic results are optimal when ζ is small enough.

Theorem 17 *Let $n = 2k-1, \zeta \leq \frac{1}{8k}, \pi_{\min} \leq \frac{1}{4k}, \varepsilon > 0$ and $\varepsilon < \min\{\frac{\pi_{\min}}{4\sqrt{k}}, \zeta\}$. Then, there exist models $(\pi, \mathbf{m}), (\pi', \mathbf{m}') \in \mathcal{D}_{n,\zeta,\pi_{\min}}$ such that $d_{\text{stat}}(\gamma(\pi,\mathbf{m}), \gamma(\pi',\mathbf{m}')) \leq (k\zeta)^{\Omega(k)}\varepsilon$, but $d_{\text{model}}((\pi,\mathbf{m}),(\pi',\mathbf{m}')) > \varepsilon$.*

Corollary 18 *For $n = 2k-1$ random variables X_i, sample complexity*

$$(1/(k\zeta))^{\Omega(k)}(1/\varepsilon)$$

is necessary to compute a model that w.h.p. satisfies (5).

Note that (assuming $\pi_{\min}$ is not smaller than $\zeta^{O(k)}$), the upper and lower sample complexity bounds in Corollary 11 and Corollary 18 match in the case that $\zeta \leq k^{-1-\delta}$ for some arbitrary small $\delta > 0$. Only when ζ comes closer to its maximal possible value of $\frac{1}{k-1}$, there is a gap between upper and lower bound. In the edge case, when $\zeta = \Theta(\frac{1}{k})$, the upper bound evaluates to $\exp(O(k \log k))$ and the lower bound evaluates to $\exp(\Omega(k))$. We remark that the lower bound can be shown to be tight for k-MixIID over the entire range of the parameter ζ.

The approach towards proving Theorem 17 is motivated by the following simple observation: Given a model $(\pi, \mathbf{m})$ with $\text{rank}(\mathbb{H}(\mathbf{m})) < k$, we can take a vector α in the kernel of $\mathbb{H}(\mathbf{m})$, and the models $(\pi + \lambda\alpha, \mathbf{m})$ for any small enough choice of λ will produce the same statistics as $(\pi, \mathbf{m})$. Now, if $\mathbf{m}$ has at least $k-1$ ζ-separated rows, then $\mathbb{H}(\mathbf{m})$ will be of rank k by Theorem 13, but it might still be close to being rank-deficient, as characterized by its k'th singular value. This intuition gives rise to the following result, which we prove in Appendix B:

Lemma 19 *Let* $n \geq 1, \varepsilon > 0$ *and* $\varepsilon < \min\{\frac{\pi_{\min}}{4\sqrt{k}}, \zeta\}$. *Suppose* $(\pi, \mathbf{m}) \in \mathcal{D}_{n,\zeta,\pi_{\min}}$ *and* $\sigma_k(\mathbb{H}(\mathbf{m})) = \sigma < \frac{1}{2}$. *Then, there exists* $\hat{\pi}$ *with* $\min_j \hat{\pi}_j \geq \frac{1}{4}\pi_{\min}$ *and such that* $d_{\text{model}}((\pi, \mathbf{m}), (\hat{\pi}, \mathbf{m})) > \varepsilon$ *but* $d_{\text{stat}}(\gamma(\pi, \mathbf{m}), \gamma(\hat{\pi}, \mathbf{m})) \leq 4k\sigma \cdot \varepsilon$.

Lemma 19 reduces the challenge of finding models with large model distance but small statistical distance to finding a model with a Hadamard extension $\mathbb{H}(\mathbf{m})$ that has a small k'th singular value. In Appendix B, we will show that the model of iid variables characterized by $\mathbf{m}_i = (0, \zeta, 2\zeta, \ldots, (k-1) \cdot \zeta)$ for all i, has a Hadamard extension $\mathbb{H}(\mathbf{m})$ with singular value $\sigma_k(\mathbb{H}(\mathbf{m})) = (k\zeta)^{\Omega(k)}$ (see Lemma 27). Combining this result with Lemma 19, we obtain Theorem 17. Corollary 18 then follows immediately after observing that the statistical distance d_{stat} between two models provides an upper bound for their total variation distance. The details of this statement can be found in Lemma 28 in Appendix B.

6. Discussion

Two larger questions remain to be addressed in this area. First, it would be of great interest to achieve a similar sample complexity as in Theorem 10 for the more general "learning" task. Second, it is open to characterize the set of models, for which identification (even with perfect statistics) is possible. There exist identifiable models $(\pi, \mathbf{m})$ where none of the rows of $\mathbf{m}$ has fully-separated entries. Recent work by Gordon and Schulman (2022) gives a sufficient condition for identification that is less restrictive (though more complicated) than ζ-separation. However, there is no known way of obtaining quantitative bounds on noise-stability (which are needed for our sample complexity analysis) in that less restrictive framework.[2] Note that some kind of separation assumption is unavoidable if we insist on $\mathcal{L}_\infty$-reconstruction of the model parameters: as one example, if there are j, j' s.t. $\mathbf{m}_{ij} = \mathbf{m}_{ij'}$ for all i, then it is impossible to determine π_j and $\pi_{j'}$. However, it might be possible to completely eliminate the separation assumption in favor of settling for reconstruction in transportation (Wasserstein) distance. This was achieved for k-MixIID (the k-MixProd problem where all observables are conditionally identically distributed) in Li et al. (2015), and improved in Fan and Li (2022). It is an open question whether these ideas can be extended to k-MixProd, with the goal being transportation-cost reconstruction of each of the rows of $\mathbf{m}$.

We remark that even though Theorem 10 and its corollaries are given for ζ-separated models, our algorithm itself also works under weaker conditions. In fact, all it needs is just one ζ-separated observable and two more disjoint sets of observables, each of which have a Hadamard extension (see below) with good condition number. These requirements can, for models in "general position," be met with as few as $2 \lg k + 1$ observables. The sample complexity and runtime of the algorithm scale singly-exponentially in the number of observables actually used. Hence, it is possible that, except for a small set of "adversarial" models, model identification can typically be achieved with much lower sample complexity and runtime than can be guaranteed for the worst case. In fact, this line of thought has already been pursued for the tensor decomposition problem to which we reduce k-MixProd. It was shown that the time complexity of the tensor decomposition problem significantly improves when each tensor component of the input tensor is perturbed by small random noise Bhaskara et al. (2014). Note that one can never solve the k-MixProd Identification problem with fewer than $\lg k$ observables. Below this critical threshold of observables, the Hadamard extensions cannot be full rank and therefore the mapping from model to statistics cannot be injective.

2. Quantification in the framework of Gordon and Schulman (2022) would likely be misguided, anyway, given that it is a complex yet not tight characterization; for example it excludes mixtures of subcubes Chen and Moitra (2019).

References

J. Banks, J. Garza-Vargas, A. Kulkarni, and N. Srivastava. Pseudospectral shattering, the sign function, and diagonalization in nearly matrix multiplication time. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 529–540, 2020. URL https://doi.org/10.1109/FOCS46700.2020.00056.

A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, page 594–603, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327107. URL https://doi.org/10.1145/2591796.2591881.

W. R. Blischke. Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*, 59(306):510–528, 1964. URL https://doi.org/10.1080/01621459.1964.10482176.

K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In *Proc. 21st Ann. Conf. on Learning Theory - COLT*, pages 9–20. Omnipress, 2008. URL http://colt2008.cs.helsinki.fi/papers/7-Chaudhuri.pdf.

S. Chen and A. Moitra. Beyond the low-degree algorithm: mixtures of subcubes and their applications. In *Proc. 51st Ann. ACM Symp. on Theory of Computing*, pages 869–880, 2019. URL https://doi.org/10.1145/3313276.3316375.

M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM J. Comput.*, 31(2):375–397, 2001. URL https://doi.org/10.1137/S0097539798342496.

R. de Prony. Essai expérimentale et analytique. *J. Écol. Polytech.*, 1(2):24–76, 1795.

J. A. Rhodes E. S. Allman, C. Matias. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 2009. URL https://doi.org/10.1214/09-AOS689.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1: 211–218, 1936. URL https://doi.org/10.1007/BF02288367.

Z. Fan and J. Li. Efficient algorithms for sparse moment problems without separation, 2022.

J. Feldman, R. O'Donnell, and R. A. Servedio. Learning mixtures of product distributions over discrete domains. *SIAM J. Comput.*, 37(5):1536–1564, 2008. URL https://doi.org/10.1137/060670705.

Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proc. 12th Ann. Conf. on Computational Learning Theory*, pages 53–62, July 1999. URL https://doi.org/10.1145/307400.307412.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 4th edition, 2013.

S. Gordon, B. Mazaheri, L. J. Schulman, and Y. Rabani. The sparse Hausdorff moment problem, with application to topic models, 2020.

S. L. Gordon and L. J. Schulman. Hadamard extensions and the identification of mixtures of product distributions. *IEEE Transactions on Information Theory*, 68(6):4085–4089, 2022. URL https://doi.org/10.1109/TIT.2022.3146630.

S. L. Gordon, B. Mazaheri, Y. Rabani, and L. J. Schulman. Source identification for mixtures of product distributions. In *Proc. 34th Ann. Conf. on Learning Theory - COLT*, volume 134 of *Proc. Machine Learning Research*, pages 2193–2216. PMLR, 2021. URL http://proceedings.mlr.press/v134/gordon21a.html.

S. L. Gordon, B. Mazaheri, Y. Rabani, and L. J. Schulman. Causal inference despite limited global confounding via mixture models. In *Proc. CLeaR*, 2023. URL www.cclear.cc/2023/AcceptedPapers.

N. Goyal, S. Vempala, and Y. Xiao. Fourier PCA and robust tensor decomposition. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, page 584–593, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327107. URL https://doi.org/10.1145/2591796.2591875.

P. Hall and X.-H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics*, 31(1):201 – 224, 2003. URL https://doi.org/10.1214/aos/1046294462.

N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, second edition, 2002. URL https://doi.org/10.1137/1.9780898718027.

Y. Hua and T. K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(5):814–824, 1990. URL https://doi.org/10.1109/29.56027.

A. Juan and E. Vidal. On the use of bernoulli mixture models for text classification. *Pattern Recognition*, 35 (12):2705–2710, 2002. ISSN 0031-3203. URL https://www.sciencedirect.com/science/article/pii/S0031320301002424. Pattern Recognition in Information Systems.

A. Juan and E. Vidal. Bernoulli mixture models for binary images. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 367–370 Vol.3, 2004. URL https://doi.org/10.1109/ICPR.2004.1334543.

M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th Ann. ACM Symp. on Theory of Computing*, pages 273–282, 1994. URL https://doi.org/10.1145/195058.195155.

Y. Kim, F. Koehler, A. Moitra, E. Mossel, and G. Ramnarayan. How many subpopulations is too many? Exponential lower bounds for inferring population histories. In L. Cowen, editor, *Int'l Conf. on Research in Computational Molecular Biology*, volume 11457 of *Lecture Notes in Computer Science*, pages 136–157. Springer, 2019. URL https://doi.org/10.1007/978-3-030-17083-7_9.

J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138, 1977. ISSN 0024-3795. URL https://doi.org/10.1016/0024-3795(77)90069-6.

S. E. Leurgans, R. T. Ross, and R. B. Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993. URL https://doi.org/10.1137/0614071.

Cheng Li, Bingyu Wang, Virgil Pavlu, and Javed Aslam. Conditional bernoulli mixtures for multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2482–2491, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/lij16.html.

J. Li, Y. Rabani, L. J. Schulman, and C. Swamy. Learning arbitrary statistical mixtures of discrete distributions. In *Proc. 47th Ann. ACM Symp. on Theory of Computing*, pages 743–752, 2015. URL https://doi.org/10.1145/2746539.2746584.

J. Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. Technical Report CSD-850021, R-43, UCLA Computer Science Department, June 1985.

J. Pearl. *Causality*. Cambridge, 2nd edition, 2009.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 06 2000. ISSN 1943-2631. URL https://doi.org/10.1093/genetics/155.2.945.

Y. Rabani, L. J. Schulman, and C. Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *Proc. 5th Conf. on Innovations in Theoretical Computer Science*, pages 207–224, 2014. URL https://doi.org/10.1145/2554797.2554818.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, second edition, 2000.

B. Tahmasebi, S. A. Motahari, and M. A. Maddah-Ali. On the identifiability of finite mixtures of finite product measures. (Also in "On the identifiability of parameters in the population stratification problem: A worst-case analysis," Proc. ISIT'18 pp. 1051-1055.), 2018. URL https://arxiv.org/abs/1807.05444.

H. Teicher. Identifiability of mixtures. *Annals of Mathematical Statistics*, 32:244–248, 1961.

H. Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung. *Mathematische Annalen*, 71(4):441–479, 1912. doi: 10.1007/BF01456804.

Appendix A. Analysis of Algorithm 1

We first state a result on the numerical accuracy of computing the SVD:

Lemma 20 (Golub and Loan (2013), section 5.4.1) *Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and precision $\varepsilon > 0$, the Golub-Kahan-Reinsch algorithm computes an approximate SVD given by $\mathbf{U} \in \mathbb{R}^{m\times m}, \mathbf{\Sigma} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$ such that*

- *$\mathbf{\Sigma}$ is a diagonal matrix;*
- *$\mathbf{U} = \mathbf{W} + \mathbf{\Delta U}$ and $\mathbf{V} = \mathbf{Z} + \mathbf{\Delta V}$, where $\mathbf{W}, \mathbf{Z}$ are unitary and $||\mathbf{\Delta U}||, ||\mathbf{\Delta V}|| < \varepsilon$;*
- *$\mathbf{W\Sigma Z}^\mathsf{T} = \mathbf{A} + \mathbf{\Delta A}$ with $||\mathbf{\Delta A}|| < \varepsilon||\mathbf{A}||$.*

To analyze Algorithm 1, let $\mathbf{d} = d_{\text{stat}}(\tilde{\mathrm{g}}, \mathrm{g}) = \max\{||\tilde{\mathbf{C}}_{ST} - \mathbf{C}_{ST}||_\infty, ||\tilde{\mathbf{C}}_{ST,1} - \mathbf{C}_{ST,1}||_\infty\}$ as in Definition (6) and let $\boldsymbol{\sigma}$ be as defined in (7). Furthermore, assume that the SVD in line 3 of the algorithm is calculated with precision $\varepsilon = \mathbf{d}$ according to the statement of Lemma 20. Then, we get the following bounds:

Lemma 21

(a) $\sigma_k(\tilde{\mathbf{C}}_{ST}) \geq \pi_{\min}\boldsymbol{\sigma}^2 - 2^k\mathbf{d}$

(b) $\sigma_k(\hat{\mathbf{C}}_{ST}) \geq \pi_{\min}\boldsymbol{\sigma}^2 - 2^{k+2}\mathbf{d}$

Proof (a) By a classical perturbation bound of Weyl Weyl (1912), we have $|\sigma_k(\tilde{\mathbf{C}}_{ST}) - \sigma_k(\mathbf{C}_{ST})| \leq ||\tilde{\mathbf{C}}_{ST} - \mathbf{C}_{ST}||$. Moreover, we know $||\tilde{\mathbf{C}}_{ST} - \mathbf{C}_{ST}|| \leq 2^k||\tilde{\mathbf{C}}_{ST} - \mathbf{C}_{ST}||_\infty \leq 2^k\mathbf{d}$. Combining this with Part 2 of Theorem 13, we get the desired result.

(b) Let $\mathbf{U\Sigma V}^\mathsf{T}$ be an approximate SVD for $\tilde{\mathbf{C}}_{ST}$ satisfying the conditions of Lemma 20 with precision ε. In particular, $\mathbf{U} = \mathbf{W} + \mathbf{\Delta U}$, $\mathbf{V} = \mathbf{Z} + \mathbf{\Delta V}$ and $\mathbf{W\Sigma Z}^\mathsf{T} = \tilde{\mathbf{C}}_{ST} + \mathbf{\Delta}\tilde{\mathbf{C}}_{ST}$. We assume that the columns of $\mathbf{\Sigma}$ are ordered by magnitude of their diagonal entries. Let $\hat{\mathbf{U}}, \hat{\mathbf{V}}, \hat{\mathbf{W}}, \hat{\mathbf{Z}}, \mathbf{\Delta}\hat{\mathbf{U}}, \mathbf{\Delta}\hat{\mathbf{V}}$ denote the first k columns of the respective matrices. We get

$$\begin{aligned}
\sigma_k(\hat{\mathbf{C}}_{ST}) &= \sigma_k(\hat{\mathbf{U}}^\mathsf{T}\tilde{\mathbf{C}}_{ST}\hat{\mathbf{V}}) = \sigma_k((\hat{\mathbf{W}} + \mathbf{\Delta}\hat{\mathbf{U}})^\mathsf{T}\tilde{\mathbf{C}}_{ST}(\hat{\mathbf{Z}} + \mathbf{\Delta}\hat{\mathbf{V}})) \\
&\geq \sigma_k(\hat{\mathbf{W}}^\mathsf{T}\tilde{\mathbf{C}}_{ST}\hat{\mathbf{Z}}) - ||(\mathbf{\Delta}\hat{\mathbf{U}})^\mathsf{T}\tilde{\mathbf{C}}_{ST}\hat{\mathbf{Z}}|| - ||\hat{\mathbf{W}}^\mathsf{T}\tilde{\mathbf{C}}_{ST}\mathbf{\Delta}\hat{\mathbf{V}}|| - ||(\mathbf{\Delta}\hat{\mathbf{U}})^\mathsf{T}\tilde{\mathbf{C}}_{ST}\mathbf{\Delta}\hat{\mathbf{V}}|| \\
&\geq \sigma_k(\hat{\mathbf{W}}^\mathsf{T}\tilde{\mathbf{C}}_{ST}\hat{\mathbf{Z}}) - 3\varepsilon||\tilde{\mathbf{C}}_{ST}|| \\
&\geq \sigma_k(\hat{\mathbf{W}}^\mathsf{T}(\tilde{\mathbf{C}}_{ST} + \mathbf{\Delta}\tilde{\mathbf{C}}_{ST})\hat{\mathbf{Z}}) - ||\hat{\mathbf{W}}^\mathsf{T}\mathbf{\Delta}\tilde{\mathbf{C}}_{ST}\hat{\mathbf{V}}|| - 3\varepsilon||\tilde{\mathbf{C}}_{ST}|| \\
&\geq \sigma_k((\tilde{\mathbf{C}}_{ST} + \mathbf{\Delta}\tilde{\mathbf{C}}_{ST})) - 4\varepsilon||\tilde{\mathbf{C}}_{ST}|| \\
&\geq \sigma_k(\tilde{\mathbf{C}}_{ST}) - 5\varepsilon||\tilde{\mathbf{C}}_{ST}||.
\end{aligned}$$

Since the entries of $\tilde{\mathbf{C}}_{ST}$ are all at most 1, we have $||\tilde{\mathbf{C}}_{ST}|| \leq 2^{k-1}$ and with $\varepsilon = \mathbf{d}$, the result follows. ■

At this point, we can view $\hat{\mathbf{C}}_{ST}$ and $\hat{\mathbf{C}}_{ST,1}$ as the two slices of a $k \times k \times 2$-tensor $\hat{\mathcal{T}}$ that is close to the tensor $\mathcal{T} = [\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]), \hat{\mathbf{V}}^\mathsf{T}\mathbb{H}(\mathbf{m}[T]), \mathbb{H}(\mathbf{m}_1) \cdot \pi_\odot]$ (with the two slices $\mathcal{T}(:,:,0) = \hat{\mathbf{U}}^\mathsf{T}\mathbf{C}_{ST}\hat{\mathbf{V}}$ and $\mathcal{T}(:,:,1) = \hat{\mathbf{U}}^\mathsf{T}\mathbf{C}_{ST,1}\hat{\mathbf{V}}$). The following lemma bounds the distance between $\hat{\mathcal{T}}$ and $\mathcal{T}$, and shows that the first two components of $\mathcal{T}$ are full rank and well-conditioned.

Lemma 22 *Suppose that $\mathbf{d} \leq \pi_{\min}\boldsymbol{\sigma}^2/(k2^{2k+2})$, then*

(a) $||\hat{\mathbf{C}}_{ST} - \hat{\mathbf{U}}^\mathsf{T}\mathbf{C}_{ST}\hat{\mathbf{V}}||_\infty, ||\hat{\mathbf{C}}_{ST,1} - \hat{\mathbf{U}}^\mathsf{T}\mathbf{C}_{ST,1}\hat{\mathbf{V}}||_\infty < 2^{2k}\mathbf{d}$

(b) $\kappa(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])), \kappa(\hat{\mathbf{V}}^\mathsf{T}\mathbb{H}(\mathbf{m}[T])) \leq k^2 2^{2k+1}/(\pi_{\min}\boldsymbol{\sigma}^2)$.

Proof (a) Remember that $\hat{\mathbf{U}} = \hat{\mathbf{W}} + \boldsymbol{\Delta}\hat{\mathbf{U}}, \hat{\mathbf{V}} = \hat{\mathbf{Z}} + \boldsymbol{\Delta}\hat{\mathbf{V}}$, where $||\boldsymbol{\Delta}\hat{\mathbf{U}}||, ||\boldsymbol{\Delta}\hat{\mathbf{V}}|| < \mathbf{d}$ and the columns of $\hat{\mathbf{W}}, \hat{\mathbf{Z}}$ are orthonormal. In particular, this implies that entries of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are bounded by $1 + k\mathbf{d} \leq 2$. Hence, we have

$$\begin{aligned}&||\hat{\mathbf{C}}_{ST} - \hat{\mathbf{U}}^\mathsf{T}\mathbf{C}_{ST}\hat{\mathbf{V}}||_\infty = ||\hat{\mathbf{U}}^\mathsf{T}(\tilde{\mathbf{C}}_{ST} - \mathbf{C}_{ST})\hat{\mathbf{V}}||_\infty \\ \leq &||\hat{\mathbf{U}}^\mathsf{T}||_\infty \cdot 2^{k-1} \cdot ||\tilde{\mathbf{C}}_{ST} - \mathbf{C}_{ST}||_\infty \cdot ||\hat{\mathbf{V}}||_\infty \cdot 2^{k-1} \leq 2^{2k}\mathbf{d}.\end{aligned}$$

The result follows analogously for $\hat{\mathbf{C}}_{ST,1}$.

(b) First, since the entries of $\mathbb{H}(\mathbf{m}[S])$ are bounded by 1, we have $\sigma_1(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])) \leq k \cdot ||\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])||_\infty \leq k2^k$, and similarly, $\sigma_1(\mathbb{H}(\mathbf{m}[T])^\mathsf{T}\hat{\mathbf{V}}) \leq k2^k$. Using part (a), Lemma 21, and the assumption on $\mathbf{d}$, we get

$$\begin{aligned}\sigma_k(\hat{\mathbf{U}}^\mathsf{T}\mathbf{C}_{ST}\hat{\mathbf{V}}) &\geq \sigma_k(\hat{\mathbf{C}}_{ST}) - ||\hat{\mathbf{C}}_{ST} - \hat{\mathbf{U}}^\mathsf{T}\mathbf{C}_{ST}\hat{\mathbf{V}}|| \geq \sigma_k(\hat{\mathbf{C}}_{ST}) - k2^{2k}\mathbf{d} \\ &\geq \pi_{\min}\boldsymbol{\sigma}^2 - 2^{k+2}\mathbf{d} - k2^{2k}\mathbf{d} \geq \pi_{\min}\boldsymbol{\sigma}^2/2\end{aligned}$$

Hence, we deduce

$$\begin{aligned}\pi_{\min}\boldsymbol{\sigma}^2/2 &\leq \sigma_k(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])\pi_\odot\mathbb{H}(\mathbf{m}[T])^\mathsf{T}\hat{\mathbf{V}}) \\ &\leq \sigma_k(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])) \cdot \sigma_1(\pi_\odot) \cdot \sigma_1(\mathbb{H}(\mathbf{m}[T])^\mathsf{T}\hat{\mathbf{V}}) \\ &\leq \sigma_k(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])) \cdot k2^k.\end{aligned}$$

We conclude that $\kappa(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])) = \sigma_1(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]))/\sigma_k(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])) \leq 2 \cdot (k2^k)^2/(\pi_{\min}\boldsymbol{\sigma}^2)$, and the result follows analogously for $\kappa(\hat{\mathbf{V}}^\mathsf{T}\mathbb{H}(\mathbf{m}[T]))$. ■

The following result from Bhaskara et al. (2014) provides us with error bounds for the core step of the tensor decomposition algorithm we are using.

Theorem 23 (Bhaskara et al. (2014), Theorem 2.3) *Let $\varepsilon > 0$ and $\mathcal{T}, \hat{\mathcal{T}}$ be two $k \times k \times 2$-tensors, such that*

- *$\mathcal{T} = [\mathbf{X}, \mathbf{Y}, \mathbf{Z}]$ with $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{k\times k}, \mathbf{Z} \in \mathbb{R}^{2\times k}$;*
- *$\kappa(\mathbf{X}), \kappa(\mathbf{Y}) \leq \kappa$;*
- *the entries of $(\mathbf{Z}_{1k}\mathbf{Z}_{2k}^{-1})_k$ are ζ-separated;*
- *$||\mathbf{X}_i||_2, ||\mathbf{Y}_i||_2, ||\mathbf{Z}_i||_2$ are bounded by a constant;*
- *$||\hat{\mathcal{T}} - \mathcal{T}||_\infty < \varepsilon \cdot poly(1/\kappa, 1/k, \zeta)$;*

then the eigenvectors of $\mathcal{T}(:,:,1)\mathcal{T}(:,:,0)^{-1}$ and $\mathcal{T}(:,:,1)^\mathsf{T}(\mathcal{T}(:,:,0)^\mathsf{T})^{-1}$ approximate the columns $\mathbf{X}_i$ and $\mathbf{Y}_i$ respectively, up to permutation and additive error ε.

(*Comment:* This is slightly more specific than the theorem statement in the reference, but it easily follows from the general statement.)

Finally, we will make use of the following classical result on perturbations of linear systems:

Lemma 24 (Higham (2002), Section 7.1) *Let $\mathbf{A}x = b$ and $(\mathbf{A} + \boldsymbol{\Delta}\mathbf{A})y = b + \Delta b$, where $||\boldsymbol{\Delta}\mathbf{A}|| \leq \gamma||\mathbf{A}||$ and $||\Delta b|| \leq \gamma||b||$, and assume that $\gamma \cdot \kappa(\mathbf{A}) < 1$. Then,*

$$\frac{||x - y||_2}{||x||_2} \leq \frac{2\gamma \cdot \kappa(\mathbf{A})}{1 - \gamma \cdot \kappa(\mathbf{A})}.$$

Now, we have all the tools to prove Theorem 10.

Proof of Theorem 10 Fix $\varepsilon \in (0, \zeta/2)$ and a model $(\pi, \mathbf{m}) \in \mathcal{D}_{2k-1,\zeta,\pi_{\min}}$ with statistics g, and suppose that Algorithm 1 is given approximate statistics $\tilde{\mathrm{g}}$ with $||\,\mathrm{g} - \tilde{\mathrm{g}}\,||_\infty = \mathbf{d} \leq \varepsilon \cdot (\pi_{\min}\zeta^k)^C$ for some large constant C. Consider the $k \times k \times 2$-tensor $\hat{\mathcal{T}}$ that consists of the slices $\hat{\mathbf{C}}_{ST}$ and $\hat{\mathbf{C}}_{ST,1}$ and $\mathcal{T} = \left[\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S]), \hat{\mathbf{V}}^\mathsf{T}\mathbb{H}(\mathbf{m}[T]), \mathbb{H}(\mathbf{m}_1)\pi_\odot\right]$. Define $\kappa = \max\{\kappa(\hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])), \kappa(\hat{\mathbf{V}}^\mathsf{T}\mathbb{H}(\mathbf{m}[T]))\}$. By Lemma 22, we have $\kappa \leq \frac{1}{\pi_{\min}} \cdot \left(\frac{1}{\zeta}\right)^{O(k)}$ and $||\hat{\mathcal{T}} - \mathcal{T}||_\infty \leq \exp(O(k))\mathbf{d}$. Hence, by Theorem 23, we get $||\hat{\mathbf{S}} - \hat{\mathbf{U}}^\mathsf{T}\mathbb{H}(\mathbf{m}[S])||_\infty, ||\hat{\mathbf{T}} - \hat{\mathbf{V}}^\mathsf{T}\mathbb{H}(\mathbf{m}[T])||_\infty \leq \exp(O(k))\mathbf{d} \cdot \mathrm{poly}(\kappa, k, 1/\zeta) = \left(\frac{1}{\pi_{\min}}\right)^{O(1)}\left(\frac{1}{\zeta}\right)^{O(k)}\mathbf{d}$, after possibly permuting the columns of $\hat{\mathbf{S}}$ and $\hat{\mathbf{T}}$. Now, $\tilde{\pi}$ is defined via a linear system that is a perturbation of equation (3). If $\mathbf{d}$ is small enough, then the conditions of Lemma 24 are satisfied with $\gamma = \varepsilon/(4\kappa)$, and we get $||\pi - \tilde{\pi}||_2 \leq \varepsilon$. Similarly, we get $\tilde{\mathbf{m}}_i$ from a perturbed version of equation (4) and we can use Lemma 24 with $\gamma = \varepsilon/(4\kappa \cdot \pi_{\min})$ to deduce $||\mathbf{m}_i - \tilde{\mathbf{m}}_i||_2 \leq \varepsilon$. Changing the norms to $||.||_\infty$ incurs at most another factor of k, by which we can decrease $\mathbf{d}$, so then the output of Algorithm 1 satisfies $d_{\mathrm{model}}((\pi, \mathbf{m}), (\tilde{\pi}, \tilde{\mathbf{m}})) < \varepsilon$.

To prove the second part of the theorem, suppose the model $(\tilde{\pi}, \tilde{\mathbf{m}})$ has statistics $\tilde{\mathrm{g}}$, and again $||\,\mathrm{g} - \tilde{\mathrm{g}}\,||_\infty \leq \varepsilon \cdot (\pi_{\min}\zeta^k)^C$. By Lemma 21, $\tilde{\mathbf{C}}_{ST}$ has rank k. Hence, by equation (1), $\tilde{\mathbf{m}}_1$ is the vector of eigenvalues of $\hat{\mathbf{C}}_{ST,1}\hat{\mathbf{C}}_{ST}^{-1}$. At the same time, by the first part of the theorem, these eigenvalues give an ε-approximation to $\mathbf{m}$ (after permuting them), and $\varepsilon < \zeta/2$ implies that they must be separated. We essentially proved in section 3.1 through equations (1) - (4) that $\tilde{\mathbf{C}}_{ST}$ having rank k and separation of $\tilde{\mathbf{m}}_1$ are sufficient conditions for Algorithm 1 to perfectly recover $(\tilde{\pi}, \tilde{\mathbf{m}})$ given perfect statistics $\tilde{\mathrm{g}} = \gamma(\tilde{\pi}, \tilde{\mathbf{m}})$. But then, the first part of the theorem implies $d_{\mathrm{model}}((\tilde{\pi}, \tilde{\mathbf{m}}), (\pi, \mathbf{m})) \leq \varepsilon$, as desired. ■

Proof of Corollary 12 Suppose we have $n \geq 2k-1$ variables and we know a subset $A \subseteq [n]$ of $2k-1$ ζ-separated variables. Then, Algorithm 1 can be used to identify $\mathbf{m}[A]$ and π up to error ε in runtime $\exp(O(k))$ using $(1/\zeta)^{O(k)}(1/\pi_{\min})^{O(1)}(1/\varepsilon)^2$ many samples. For each $i \notin A$, we can then compute

$$\tilde{\mathbf{m}}_i = ((\tilde{\mathrm{g}}(R \cup \{i\}))_{R \subseteq T})^\mathsf{T} \cdot \hat{\mathbf{V}} \cdot \left(\hat{\mathbf{T}}^\mathsf{T}\right)^{-1} \cdot \tilde{\pi}_\odot^{-1}$$

as in line 9 of the algorithm (where $T \subseteq A$ is a set of size $k-1$ and $\hat{\mathbf{T}}$ approximates $\hat{\mathbf{V}}^\mathsf{T}\mathbb{H}(\mathbf{m}[T])$). This takes runtime at most $n \cdot \exp(O(k))$. Given that $||(\tilde{\mathrm{g}}(R \cup \{i\}))_{R \subseteq T} - (\mathrm{g}(R \cup \{i\}))_{R \subseteq T}||_\infty \leq \varepsilon(\pi_{\min}\zeta^k)^C$ for some large enough C, the same analysis as for Theorem 10 asserts that $\tilde{\mathbf{m}}_i$ is an ε-approximation to $\mathbf{m}_i$ (here, g is the vector of perfect statistics). Hence, computing ε-approximations to $\mathbf{m}_i$ for all $i \notin A$ requires obtaining $(\varepsilon(\pi_{\min}\zeta^k)^C)$-approximations to $(n - |A|) \cdot 2^{k-1}$ entries of the observable moment vector g. Standard Chernoff bounds and a union bound show that this is possible with $\log n \cdot (1/\zeta)^{O(k)}(1/\pi_{\min})^{O(1)}(1/\varepsilon)^2$ many samples.

If the subset of $2k-1$ ζ-separated variables is not known (but guaranteed to exist), we can simply run Algorithm 1 to identify $\mathbf{m}[A]$ for all $\binom{n}{2k-1}$ possible guesses A of this subset (stopping the algorithm when encountering any issues that might occur when $\hat{\mathbf{C}}_{ST}$ is not invertible). Then, we compute the vector of statistics $\gamma(\tilde{\pi}, \tilde{\mathbf{m}}[A]) = \mathbb{H}(\tilde{\mathbf{m}}[A])\tilde{\pi}$ for each valid output $(\tilde{\pi}, \tilde{\mathbf{m}}[A])$ and choose the model $(\tilde{\pi}, \tilde{\mathbf{m}}[A])$ that minimizes the distance $\mathbf{d}(A) = ||\gamma(\tilde{\pi}, \tilde{\mathbf{m}}[A]) - \tilde{\mathrm{g}}[2^A]||_\infty$. All this takes runtime at most $n^{2k}\exp(O(k))$. Let A^* be the correct guess and suppose we have $||\,\tilde{\mathrm{g}}[2^{A^*}] - \mathrm{g}[2^{A^*}]||_\infty \leq \delta$. Then, by Theorem 10, we get $d_{\mathrm{model}}((\tilde{\pi}, \tilde{\mathbf{m}}[A^*]), (\pi, \mathbf{m}[A^*])) \leq \delta \cdot (\pi_{\min}\zeta^k)^{-C}$ for some large, positive constant C. Hence, we have (after possibly permuting $\tilde{\pi}, \tilde{\mathbf{m}}[A^*]$) that $||\pi - \tilde{\pi}||_\infty \leq \delta \cdot (\pi_{\min}\zeta^k)^{-C}$ and $||\mathbb{H}(\tilde{\mathbf{m}}[A^*]) - \mathbb{H}(\mathbf{m}[A^*])||_\infty \leq$

$2^{2k}\delta \cdot (\pi_{\min}\zeta^k)^{-C}$. This implies

$$\begin{aligned}
\mathbf{d}(A^*) &= ||\gamma(\tilde{\pi}, \tilde{\mathbf{m}}[A^*]) - \tilde{\mathrm{g}}[2^{A^*}]||_\infty \\
&\leq ||\mathbb{H}(\tilde{\mathbf{m}}[A^*])\tilde{\pi} - \mathbb{H}(\mathbf{m}[A^*])\pi||_\infty + ||\,\mathrm{g}[2^{A^*}] - \tilde{\mathrm{g}}[2^{A^*}]||_\infty \\
&\leq ||\mathbb{H}(\tilde{\mathbf{m}}[A^*])\tilde{\pi} - \mathbb{H}(\mathbf{m}[A^*])\tilde{\pi}||_2 + ||\mathbb{H}(\mathbf{m}[A^*])\tilde{\pi} - \mathbb{H}(\mathbf{m}[A^*])\pi||_2 + \delta \\
&\leq ||\mathbb{H}(\tilde{\mathbf{m}}[A^*]) - \mathbb{H}(\mathbf{m}[A^*])||_2 \cdot ||\tilde{\pi}||_2 + ||\mathbb{H}(\mathbf{m}[A^*])||_2 \cdot ||\tilde{\pi} - \pi||_2 + \delta \\
&\leq 2^k||\mathbb{H}(\tilde{\mathbf{m}}[A^*]) - \mathbb{H}(\mathbf{m}[A^*])||_\infty \cdot \sqrt{k} + 2^k \cdot \sqrt{k}||\tilde{\pi} - \pi||_\infty + \delta \\
&\leq 2^{4k}\delta \cdot (\pi_{\min}\zeta^k)^{-C}.
\end{aligned}$$

If $\delta < 2^{-4k}(\pi_{\min}\zeta^k)^{2C}\varepsilon$, then the minimal distance $\mathbf{d}(A)$ is at most $\varepsilon(\pi_{\min}\zeta^k)^C$. Hence, by Theorem 10, the model $(\tilde{\pi}, \tilde{\mathbf{m}}[A])$ is an ε-approximation of $(\pi, \mathbf{m}[A])$ and we can proceed as in the first part of the proof to get a full ε-approximation to $(\pi, \mathbf{m})$. Ensuring that $||\,\tilde{\mathrm{g}}[2^A] - \mathrm{g}[2^A]||_\infty \leq 2^{-4k}(\pi_{\min}\zeta^k)^{2C}\varepsilon$ for all subsets of $A \subseteq [n]$ of size $2k-1$ requires $\log(n^{2k}) \cdot 2^{8k}(\pi_{\min}\zeta^k)^{-4C}\varepsilon^{-2} = \log n \cdot (1/\pi_{\min})^{O(1)}(1/\zeta)^{O(k)}(1/\varepsilon)^2$ many samples. ■

Appendix B. Proof of the lower bound

Proof of Lemma 19 Let $(\pi, \mathbf{m}) \in \mathcal{D}_{n,\zeta,\pi_{\min}}$ and $\sigma_k(\mathbb{H}(\mathbf{m})) = \sigma < \frac{1}{2}$. Let $\mathbb{H}(\tilde{\mathbf{m}})$ be the best rank-$(k-1)$-approximation of $\mathbb{H}(\mathbf{m})$. By the Eckart-Young Theorem Eckart and Young (1936), we have $||\mathbb{H}(\tilde{\mathbf{m}}) - \mathbb{H}(\mathbf{m})|| = \sigma$. Let $\alpha \in \mathbb{R}^k$ be a vector in the right kernel of $\mathbb{H}(\tilde{\mathbf{m}})$ with $||\alpha||_2 = 1$. Let $\mathbb{1}$ denote the all-ones vector in $\mathbb{R}^k$ and let e_1 denote the vector whose first entry is 1 and all other entries are zero. We have

$$|\mathbb{1}^\mathsf{T}\alpha| = |(\mathbb{H}(\mathbf{m})\alpha)_1| = |((\mathbb{H}(\mathbf{m}) - \mathbb{H}(\tilde{\mathbf{m}}))\alpha)_1| \leq ||\mathbb{H}(\mathbf{m}) - \mathbb{H}(\tilde{\mathbf{m}})|| \cdot ||\alpha||_2 \leq \sigma.$$

Now, define $\hat{\pi} = \pi + 2\sqrt{k}\varepsilon \cdot (\alpha - (\mathbb{1}^\mathsf{T}\alpha)e_1)$. First, we check that $\hat{\pi}$ is a valid probability vector. By our assumptions, we have

$$||2\sqrt{k}\varepsilon \cdot (\alpha - (\mathbb{1}^\mathsf{T}\alpha)e_1)||_\infty \leq 2\sqrt{k}\varepsilon \cdot (||\alpha||_\infty + |\mathbb{1}^\mathsf{T}\alpha|) < \frac{\pi_{\min}}{2} \cdot (1+\sigma) \leq \frac{3}{4}\pi_{\min},$$

hence, all the entries of $\hat{\pi}$ are larger than $\frac{1}{4}\pi_{\min}$. Moreover, we have

$$\sum_{j=1}^{k} \hat{\pi}_j = \mathbb{1}^\mathsf{T}\pi + 2\sqrt{k}\varepsilon \cdot (\mathbb{1}^\mathsf{T}\alpha - \mathbb{1}^\mathsf{T}\alpha) = 0.$$

Now, recall the definition of d_{model} as

$$d_{\mathrm{model}}((\pi, \mathbf{m}), (\hat{\pi}, \mathbf{m})) = \min_{\rho \in S_k} \max\{\max_j |\pi_j - \hat{\pi}_{\rho(j)}|, \max_{i,j} |\mathbf{m}_{i,j} - \mathbf{m}_{i,\rho(j)}|\}.$$

For any permutation ρ that is not the identity, we have $\max_{i,j} |\mathbf{m}_{i,j} - \mathbf{m}_{i,\rho(j)}| \geq \zeta > \varepsilon$ by ζ-separation of $\mathbf{m}$. For the identity permutation, we have

$$\max_j |\pi_j - \hat{\pi}_j| = ||2\sqrt{k}\varepsilon \cdot (\alpha - (\mathbb{1}^\mathsf{T}\alpha)e_1)||_\infty \geq 2\varepsilon \cdot (||\alpha||_2 - ||(\mathbb{1}^\mathsf{T}\alpha)e_1||_2) \geq 2\varepsilon \cdot (1-\sigma) > \varepsilon,$$

so we conclude $d_{\text{model}}((\pi, \mathbf{m}), (\hat{\pi}, \mathbf{m})) > \varepsilon$. Moreover, we have

$$\begin{aligned} d_{\text{stat}}(\gamma(\pi, \mathbf{m}), \gamma(\hat{\pi}, \mathbf{m})) &= ||\mathbb{H}(\mathbf{m})\pi - \mathbb{H}(\mathbf{m})\hat{\pi}||_\infty \\ &= 2\sqrt{k}\varepsilon \cdot ||\mathbb{H}(\mathbf{m})(\alpha - (\mathbb{1}^\mathsf{T}\alpha)e_1)||_\infty \\ &\leq 2\sqrt{k}\varepsilon \cdot \left(||\mathbb{H}(\mathbf{m})\alpha||_\infty + |\mathbb{1}^\mathsf{T}\alpha| \cdot ||\mathbb{H}(\mathbf{m})||_\infty\right) \\ &\leq 2\sqrt{k}\varepsilon \cdot \left(\sqrt{k} \cdot ||(\mathbb{H}(\mathbf{m}) - \mathbb{H}(\tilde{\mathbf{m}}))\alpha||_2 + \sigma\right) \\ &\leq 2\sqrt{k}\varepsilon \cdot \left(\sqrt{k}\sigma + \sigma\right) \leq 4k\sigma \cdot \varepsilon. \end{aligned}$$

■

To find a model $(\pi, \mathbf{m})$ that has a Hadamard extension with small k'th singular value, it will be sufficient to consider a model with iid variables, i.e. a model, for which all the rows of $\mathbf{m}$ are the same. In this case, the Hadamard extension will resemble a Vandermonde matrix.

Definition 25 (Vandermonde matrix) *The Vandermonde matrix* $\mathrm{Vdm}(m) \in \mathbb{R}^{k\times k}$ *associated with a row vector* $m \in \mathbb{R}^k$ *has entries* $\mathrm{Vdm}(m)_{ij} = (m_j)^i$ *for* $i \in \{0, 1, \ldots, k-1\}$ *and* $j \in \{1, 2, \ldots, k\}$. *We also write* $\mathrm{Vdm}(m, r)$ *for the* $r \times k$ *matrix with entries* $\mathrm{Vdm}(m, r)_{ij} = (m_j)^i$.

The condition number of Vandermonde matrices is well studied, see the following result.

Lemma 26 (Higham (2002), section 22.1) *For any row vector* $m \in \mathbb{R}^k$, *we have*

$$||\mathrm{Vdm}(m)^{-1}||_\infty \geq \max_i \prod_{j\neq i} \frac{\max\{1, |m_j|\}}{|m_i - m_j|}.$$

Lemma 27 *Let* $n \geq k-1$. *For any* $\zeta \leq \frac{1}{k}$, *there exists a matrix* $\mathbf{m} \in [0,1]^{n\times k}$ *with* ζ*-separated columns and* $\sigma_k(\mathbb{H}(\mathbf{m})) \leq n2^n \cdot (k\zeta)^k$.

Proof Given ζ, define $\mathbf{m}$ as the matrix with n identical rows of the form $\mathbf{m}_i = (0, \zeta, 2\zeta, \ldots, (k-1)\cdot\zeta)$. According to Lemma 26, we have

$$\sigma_k(\mathrm{Vdm}(\mathbf{m}_1)) = ||\mathrm{Vdm}(\mathbf{m}_1)^{-1}||^{-1} \leq k||\mathrm{Vdm}(\mathbf{m}_1)^{-1}||_\infty^{-1} \leq k \cdot (k-1)!\zeta^{k-1} = k!\zeta^{k-1}.$$

Now, consider the matrix $\mathrm{Vdm}(\mathbf{m}_1, n) = \binom{\mathrm{Vdm}(\mathbf{m}_1)}{\mathbf{R}}$, where $\mathbf{R} \in \mathbb{R}^{(n-k)\times k}$ denotes the last $n-k$ columns of $\mathrm{Vdm}(\mathbf{m}_1, n)$. Note that $||\mathbf{R}|| \leq n \cdot ||\mathbf{R}||_\infty \leq n \cdot ((k-1)\cdot\zeta)^k$. Let $q \in \mathbb{R}^k, ||q||_2 = 1$ such that $||\mathrm{Vdm}(\mathbf{m}_1)q||_2$ is minimized. We have

$$||\mathrm{Vdm}(\mathbf{m}_1, n)q||_2 \leq ||\mathrm{Vdm}(\mathbf{m}_1)q||_2 + ||\mathbf{R}q||_2 \leq \sigma_k(\mathrm{Vdm}(\mathbf{m}_1)) + ||\mathbf{R}|| \leq n \cdot k^k \cdot \zeta^k,$$

hence, $\sigma_k(\mathrm{Vdm}(\mathbf{m}_1, n)) \leq n \cdot k^k \cdot \zeta^k$. Finally, note that $\mathbb{H}(\mathbf{m})$ is a matrix with 2^n rows that are all duplicates of some row in $\mathrm{Vdm}(\mathbf{m}_1, n)$. Hence, we have $\sigma_k(\mathbb{H}(\mathbf{m})) \leq n2^n \cdot (k\zeta)^k$. ■

This gives all we need to prove Theorem 17.

Proof of Theorem 17 Fix $n, \zeta, \pi_{\min}$, and ε as in the statement of Theorem 17. Let $\pi := (1/k, \ldots, 1/k)^\mathsf{T}$. According to Lemma 27, there exists $\mathbf{m}$ such that $(\pi, \mathbf{m}) \in \mathcal{D}_{n,\zeta,\pi_{\min}}$ and $\sigma_k(\mathbb{H}(\mathbf{m})) \leq n2^n \cdot (k\zeta)^k = (2k-1)2^{2k-1} \cdot (k\zeta)^k < \frac{1}{2}$. Now, by Lemma 19, there exists $\hat{\pi}$ such that $(\hat{\pi}, \mathbf{m}) \in \mathcal{D}_{n,\zeta,\pi_{\min}}, d_{\text{model}}((\pi, \mathbf{m}), (\hat{\pi}, \mathbf{m})) > \varepsilon$ and $d_{\text{stat}}(\gamma(\pi, \mathbf{m}), \gamma(\hat{\pi}, \mathbf{m})) \leq 4k\sigma_k(\mathbb{H}(\mathbf{m})) \cdot \varepsilon \leq (k\zeta)^{\Omega(k)} \cdot \varepsilon$. ■

Corollary 18 then follows with the following Lemma:

Lemma 28 *Let $(\pi, \mathbf{m})$ and $(\pi', \mathbf{m}')$ be two mixture models with n variables and k latent states. Then,*

$$d_{TV}((\pi, \mathbf{m}), (\pi', \mathbf{m}')) \leq 2^{2n} \cdot d_{\mathrm{stat}}((\pi, \mathbf{m}), (\pi', \mathbf{m}')).$$

Proof The models $(\pi, \mathbf{m})$ and $(\pi, \mathbf{m}')$ define two different probability measures on $\{0,1\}^n$, which we denote by μ and μ'. We calculate

$$\begin{aligned}
d_{TV}((\pi, \mathbf{m}), (\pi', \mathbf{m}')) &= \frac{1}{2} \sum_{v \in \{0,1\}^n} \left|\mu(v) - \mu'(v)\right| \\
&= \frac{1}{2} \sum_{v \in \{0,1\}^n} \left| \sum_{j=1}^{k} \pi_j \prod_{i:v_i=1} \mathbf{m}_{ij} \prod_{i:v_i=0} (1 - \mathbf{m}_{ij}) - \sum_{j=1}^{k} \pi'_j \prod_{i:v_i=1} \mathbf{m}'_{ij} \prod_{i:v_i=0} (1 - \mathbf{m}'_{ij}) \right| \\
&= \frac{1}{2} \sum_{v \in \{0,1\}^n} \left| \sum_{S \subseteq \{i:v_i=0\}} (-1)^{|S|} \left(\sum_{j=1}^{k} \pi_j \prod_{i:v_i=1} \mathbf{m}_{ij} \prod_{i \in S} \mathbf{m}_{ij} - \sum_{j=1}^{k} \pi'_j \prod_{i:v_i=1} \mathbf{m}'_{ij} \prod_{i \in S} \mathbf{m}'_{ij} \right) \right| \\
&\leq \frac{1}{2} \sum_{v \in \{0,1\}^n} \sum_{S \subseteq \{i:v_i=0\}} \left| \sum_{j=1}^{k} \pi_j \prod_{i:v_i=1} \mathbf{m}_{ij} \prod_{i \in S} \mathbf{m}_{ij} - \sum_{j=1}^{k} \pi'_j \prod_{i:v_i=1} \mathbf{m}'_{ij} \prod_{i \in S} \mathbf{m}'_{ij} \right| \\
&\leq \frac{1}{2} \sum_{v \in \{0,1\}^n} \sum_{S \subseteq \{i:v_i=0\}} d_{\mathrm{stat}}((\pi, \mathbf{m}), (\pi', \mathbf{m}')) \leq 2^{2n} \cdot d_{\mathrm{stat}}((\pi, \mathbf{m}), (\pi', \mathbf{m}')).
\end{aligned}$$

■