The Identification of Discrete Mixture Models

Thesis by
Spencer Lane Gordon

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Computer Science

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended January 19, 2023

Spencer Lane Gordon
ORCID: 0000-0002-7101-2370

ACKNOWLEDGEMENTS

[Intentionally left blank]

ABSTRACT

In this thesis we discuss a variety of results on learning and identifying discrete mixture models, i.e., distributions that are a convex combination of k from a known class C of distributions. We first consider the case where C is the class of binomial distributions, before generalizing to the case of product distributions. We provide a necessary condition for identifiability of mixture of products distributions as well as a generalization to structured mixtures over multiple latent variables.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] S. L. Gordon and L. J. Schulman, "Hadamard extensions and the identification of mixtures of product distributions," *IEEE Transactions on Information Theory*, vol. 68, no. 6, pp. 4085–4089, 2022. DOI: 10.1109/TIT.2022.3146630,
S.G. participated in all parts of the project.

[2] S. Gordon, B. H. Mazaheri, Y. Rabani, and L. Schulman, "Source identification for mixtures of product distributions," in *Proceedings of Thirty Fourth Conference on Learning Theory*, M. Belkin and S. Kpotufe, Eds., ser. Proceedings of Machine Learning Research, vol. 134, PMLR, Aug. 2021, pp. 2193–2216. [Online]. Available: https://proceedings.mlr.press/v134/gordon21a.html,
S.G. participated in all parts of the project.

[3] S. Gordon, B. Mazaheri, L. J. Schulman, and Y. Rabani, "The sparse hausdorff moment problem, with application to topic models," *CoRR*, vol. abs/2007.08101, 2020. arXiv: 2007.08101. [Online]. Available: https://arxiv.org/abs/2007.08101,
S.G. participated in all parts of the project.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

Chapter 1

INTRODUCTION

In this thesis, we will investigate problems of the following form: We have a distribution over vectors $X = (X_1, X_2, \ldots, X_n) \in [B]^n$ for $B \in \mathbb{N}$ with $P(X_1, \ldots, X_n)$ a mixture of k distributions from a class $\mathcal{D}$ of distributions with a given parameterization. We are given samples of X and are asked to determine the mixture weights, that is, the weight of each mixture constituent in the distribution, along with the parameters determining the mixture constituents. Our goal is to recover these parameter up to an input accuracy ε with probability $1 - \delta$, where δ is also part of our input. We'd like to use as few samples as possible.

We will write our target distribution as a convex combination of k distributions $P_1(\theta_1), \ldots, P_k(\theta_k) \in \mathcal{D}$, with non-negative weights $\pi_1, \ldots, \pi_k$ summing to 1.

Definition 1. The $(k-1)$-simplex, $\Delta_{k-1} \subseteq \mathbb{R}^k$, consists of all vectors $\pi \in \mathbb{R}^k_{\geq 0}$ such that $\sum_i \pi_i = 1$.

Definition 2 (Mixture of discrete distributions)**.** A k-mixture of distributions is parameterized by distributions $P_1, \ldots, P_k$ and mixing weights $\pi \in \Delta_{k-1}$ and is the marginal distribution over $X = (X_1, X_2, \ldots, X_n)$ of the joint distribution on (X, U) given by $P(X = x \mid U = u) = P_u(X = x)$ and $P(U = u) = \pi_u$ for all $u \in [k]$.

Notation 3. When $\mathcal{D}$ is a family of discrete distributions parameterized by $\theta \in \mathbb{R}^n$ we can write a mixture of k distributions from $\mathcal{D}$ as $P(X) = \pi_1 P(X; \theta_1) + \cdots + \pi_k P(X; \theta_k)$. When we don't care about the parameterization, we will write $P(X) = \pi_1 P_1(X) + \cdots + \pi_k P_k(X)$.

The k-Mix $\mathcal{D}$ problem

- **Input:** The input to this problem is a distribution $P(X)$ that is a mixture of k distributions from $\mathcal{D}$ with unknown mixing weights $\pi_1, \ldots, \pi_k$ where the ith mixture constituent is parameterized by unknown parameter θ_i, along with parameters $\varepsilon, \delta > 0$.

- **Goal:** The goal is to design an algorithm that uses S samples from P and with probability at least $1-\delta$ outputs the following: a vector $\tilde{\pi} \in \Delta_{k-1}$ along with distributions $P(\cdot;\tilde{\theta}_1),\ldots,P(\cdot;\tilde{\theta}_k) \in \mathcal{D}$ such that $\|\tilde{\pi}-\pi\|_\infty \leq \varepsilon$ and $\left\|\tilde{\theta}_i-\theta_i\right\|_\infty \leq \varepsilon$ for $i=1,\ldots,k$.

Identifiability

There is one way in which the k-Mix $\mathcal{D}$ problem is too hard, no matter which class of distributions $\mathcal{D}$ we are considering; permuting the mixture components by applying some permutation $\sigma : [k] \to [k]$ will not change any of the observed statistics. In particular, let $P'(X) = \pi'_1 P(X;\theta'_1) + \pi'_2 P(X;\theta'_2) + \cdots + \pi'_k P(X;\theta'_k)$ where $\pi'_u = \pi_{\sigma^{-1}(u)}$ and $\theta'_u = \theta'_{\sigma^{-1}(u)}$ for all u. Then $P = P'$ and there is no way to distinguish between the distribution on X induced by these two distinct mixture models. Thus, we will always be interested in approximately recovering the parameters *only up to the equivalence class of mixture models induced by permuting mixture components*.

- Revised Goal: Our revised goal is to design an algorithm that uses S samples from P and with probability at least $1-\delta$ outputs the following: a vector $\tilde{\pi} \in \Delta_{k-1}$ along with distributions $P(\cdot;\tilde{\theta}_1),\ldots,P(\cdot;\tilde{\theta}_k) \in \mathcal{D}$ such that *for some permutation* $\sigma : [k] \to [k]$ the vector π_σ obtained by permuting coordinates according to σ satisfies $\|\tilde{\pi}-\pi_\sigma\|_\infty \leq \varepsilon$ and $\left\|\tilde{\theta}_i-\theta_{\sigma(i)}\right\|_\infty \leq \varepsilon$ for $i=1,\ldots,k$.

Having weakened our requirement for a solution to the problem, there is another way in which the k-Mix $\mathcal{D}$ problem may still be too hard for certain classes of distributions $\mathcal{D}$ and certain concrete choices of distributions from $\mathcal{D}$. There may be multiple equivalence classes of parameters that generate the same distribution on X.

To illustrate this phenomenon, consider the following class of distributions, where $n \in \mathbb{N}$ is fixed:

$$\mathrm{IID} := \{\mathrm{IID}(\theta) : \theta \in [0,1]\},$$

where $X \in \{0,1\}^n$ and $P(X=x) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$ for $P = \mathrm{IID}(\theta)$.

Example 4. In both of the following, let $\mathcal{D} = \mathrm{IID}$ and let $k = 3$.

Degenerate mixture Consider $\pi = (0, 1/2, 1/2)$ and $\theta = (\theta_1, \theta_2, \theta_3) = (1/2, 1/4, 3/4)$. Then

$$\begin{aligned} P(X = x) &= 0 \times P(X = x; \theta_1) + \frac{1}{2} \times P(X = x; 1/4) + \frac{1}{2} \times P(X = x; 3/4) \\ &= \frac{1}{2} \times P(X = x; 1/4) + \frac{1}{2} \times P(X = x; 3/4), \end{aligned}$$

so any choice of θ_1 gives the same distribution on X.

Identical mixture components Consider $\pi = (1/3, 1/3, 1/3)$ and $\theta = (1/2, 1/2, 1/4)$. Then

$$\begin{aligned} P(X = x) &= \pi_1 \times P(X = x; 1/2) + \pi_2 \times P(X = x; 1/2) + \pi_3 \times P(X = x; 1/4) \\ &= (\pi_1 + \pi_2) \times P(X = x; 1/2) + \pi_3 \times P(X = x; 1/4), \end{aligned}$$

so any choice of π_1 and π_2 that retains $\pi_1 + \pi_2 = 2/3$ will give the same distribution on X.

What we need for the k-Mix $\mathcal{D}$ problem to be solvable is *identifiability* of the mixture model.

Definition 5 (Identifiability)**.** We will say that a k-Mix $\mathcal{D}$ instance π, θ is *identifiable* when $P(X; \pi, \theta) \neq P(X; \pi', \theta')$ for any π', θ' not in the equivalence class of parameters containing π, θ.

Note that identifiability is a property of instances and not of problems. In this work we will be very interested in understanding when a given instance is identifiable.

1.1 Organization

The k-Mix IID Problem

The first case, considered in Chapter 2, will be mixtures of IID distributions, i.e., $\mathcal{D} := \text{IID}$. For the k-Mix IID problem, there is a complete characterization of identifiability in terms of conditions on the parameters and the number of copies of X_1 available. Moreover, there are known algorithms for solving the problem. In this chapter, we present a new analysis of Prony's method, a classical method for solving this problem, and obtain the best known upper bounds on the required sample complexity and runtime for solving the k-Mix IID problem. Finally, we present an algorithm for the non-binary case, where $X \in [B]^n$ for some $B \in \mathbb{N}$.

The k-Mix Prod Problem

The next case, considered in Chapters 3 and 4, is that of *product* distributions. That is,

$$P_u(X = x; \theta) = \prod_{i=1}^{n} \theta_i^{x_i}(1 - \theta_i)^{1-x_i}$$

for $\theta \in [0, 1]^n$. In Chapter 3 we present a new sufficient condition for identifiability that generalizes the identifiability condition for the k-Mix IID problem and in Chapter 4 we give an algorithm to solve the k-Mix Product problem under a stronger identifiability assumption.

Generalizations

In Chapter 5, we consider a slightly different model and prove an identifiability result for this model.

Chapter 2

THE k-MIX IID PROBLEM

[1] S. Gordon, B. Mazaheri, L. J. Schulman, and Y. Rabani, "The sparse hausdorff moment problem, with application to topic models," *CoRR*, vol. abs/2007.08101, 2020. arXiv: 2007.08101. [Online]. Available: https://arxiv.org/abs/2007.08101,

2.1 Introduction

We consider the problem of learning a mixture of k IID distributions, k-Mix IID. This is equivalent to learning a mixture of k binomial distributions; motivated by the following analogy, we call this *the k-coin problem*. (In the literature this is also called the k-spike problem.)

Consider a set of k visually indistinguishable coins. When tossed, coin j has (unknown) probability θ_j of coming up heads. There is an (unknown) probability distribution π on the coins.

Our sampling regimen is this: a coin is picked according to π. We do not discover which coin we picked. We then toss this coin m times, yielding a sequence $X = (X_1, \ldots, X_m) \in \{0,1\}^m$. Repeat.

The samples X are distributed according to

$$\Pr(X = x) = \Pr(X_1 = x_1, \ldots, X_m = x_m) = \sum_{j=1}^{k} \pi_j \prod_{i=1}^{m} \theta_j^{x_i} (1-\theta_j)^{1-x_i}. \quad (2.1)$$

We write $\mathcal{M} = (\theta, \pi)$ for the parameters of this model. Our goal is to start from empirical statistics on X which are close to the probabilities in (2.1), and from that to produce reconstructed parameters $\tilde{\pi}_1, \ldots, \tilde{\pi}_k$ and $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ such that $\left\|\theta - \tilde{\theta}\right\|_\infty$ and $\|\pi - \tilde{\pi}\|_\infty$ are small.

Equivalent formulation. The sample statistics are of course invariant under permutation of $[m]$, so only $\sum X_i$ matters. Therefore we can effectively think of the distribution in (2.1) as a mixture of binomial distributions: $\mathrm{Binomial}(m, \theta_j)$ with mixture weights (π_j).

We will show below that estimates of the probabilities in (2.1) give estimates of the following quantities,

$$\mu_i := \sum_j \pi_j \theta_j^i, \quad i = 0, \ldots, 2k, \tag{2.2}$$

which we will refer to as the *moments of the distribution* $\mathcal{P}$, where we define $\mathcal{P}$ as the probability measure on $[0, 1]$ supported on k discrete points θ_j, each with weight π_j. That is, we can write

$$\mu_i := \int_0^1 \theta^i \, \mathrm{d}\mathcal{P}(\theta) \qquad (i \geq 0). \tag{2.3}$$

Obviously as $m \to \infty$ we are able to learn from a single sample the bias of the chosen coin; and thus can determine the distribution on coins with enough samples. On the other hand, even with exact statistics, we need $m \geq 2k$ in order to verify that $\mathcal{P}$ is supported on at most k points (this will be explained below). In this chapter we will solve the problem for $m = 2k$.

Understanding the relationship between the sequence (μ_i) and measures on $[0, 1]$ is a classical problem in probability, known as the Hausdorff moment problem.

The Sparse Hausdorff Moment Problem. The Hausdorff moment problem is that of determining what moment sequences (μ_i) are possible for a probability measure $\mathcal{P}$ supported on $[0, 1]$. See [1], [2] for background on this classical problem.

Our problem is the computational version of the Hausdorff moment problem, "sparse" because the measure $\mathcal{P}$ is supported on an unknown but discrete set $\{\theta_1, \ldots, \theta_k\}$.

Motivation. The problem of reconstructing the parameters (θ, π) is quite useful:

(i) By a known reduction, any algorithm for this problem lifts to an algorithm for learning Topic Models. This will be discussed in Sec. 2.6.

(ii) The moments we analyze correspond to the cumulative distribution function of an exponential distribution. Thus, the problem of learning mixtures of binomial distributions is equivalent to learning mixtures of exponential

distributions. [3] show that learning population histories from coalescence times reduces to learning mixtures of exponential distributions. The use of [3] of the Matrix Pencil Method is interchangeable with our version of Prony's method, which improves time complexity. Other applications of learning mixtures of binomial or (equivalently) exponential distributions abound in various areas, including ecology, geology, social sciences, and computer systems. For example, see [4] for an application in network evaluation.

(iii) This problem is a special case of the problem of identifying mixture models of k product distributions on binary variables, a problem on which there has been an impressive sequence of contributions in the last two decades, as we will discuss below. In chapter 4, our results for the iid case are used to improve the complexity of the k-Mix Prod problem. In particular, we use the algorithm for k-Mix IID as a subroutine in the solution to k-Mix Prod.

(iv) Algorithms for identifying mixtures of product distributions are the simplest case of the yet-more-general problem of identifying distributions on directed graphical models [5]. There has been little work in this direction, [6] being a notable exception. However, even that work has to make strong assumptions about the distributions of the variables X_i, and in particular they cannot be binary. (Except for the case $k = 2$, but we are concerned here with complexity of the problem as a function of k.) Source identification in causal graphical models is an important direction for future research, and the dependence of the sample size and runtime complexity on k matters a great deal. (One may think of k as a bound on how much "confounding" there is among the observables in the model.)

Prony's method. The algorithm we analyze is essentially that of Prony, 1795 [7]. The idea is to (a) characterize the coin biases (the support of $\mathcal{P}$) as the roots of a polynomial whose coefficient vector is the kernel of the Hankel matrix; (b) use polynomial root-finding to determine the empirical coin biases; (c) reconstruct the mixture weights by polynomial interpolation.

Prior work on the sparse Hausdorff, i.e., k-coin mixture, problem. It has long been acknowledged in the numerical analysis literature (e.g., [1] §9.4, [8]) that the Prony method is highly sensitive to sample error (i.e., to errors in the moments). This instability is also inherent to our *problem* (the source identification of k-coin mixture models); a lower bound ([9] Thm 6.1)

shows that even for any constant $c \geq 2$, if ck (rather than just the minimum $2k$) noisy moments are available, accurate source identification is possible only if those moments are available to accuracy $\exp(-\Omega(k))$. To be specific here, the input to our problem is an empirical moment sequence $\tilde{\mu}_0, \ldots, \tilde{\mu}_m$; its accuracy is $\max_i |tilde\mu_i - \mu_i|$ where μ_i is as in (2.3).

In order to obtain this short list of moments to within accuracy ε one needs sample size roughly $1/\varepsilon^2$. Throughout the paper, therefore, accuracies of $\exp(-\Omega(k))$ in the moments, translate to sample size bounds of $\exp(\Omega(k))$, and vice versa; we shall therefore have no need to comment separately on sample size.[1]

The prior upper bounds for the problem were: (i) [9] re-invented the Prony method and solved the problem using moment accuracy $\min\{\zeta^{O(k)}, k^{-O(k^2)}\}$ and runtime $\mathrm{poly}(k)$. (ii) A different algorithm in [10] improved requirements in the moment accuracy to $k^{-O(k)}$ but required runtime $k^{O(k^2)}$. (iii) Motivated by a problem in population genetics that reduces to the k-coin mixture problem, [3] analyzed a solution using the Matrix Pencil Method (MPM), which required moment accuracy $\pi_{\min}^4 \zeta^{O(k)}$.

We note that [9], [10] do not depend on $\pi_{\min}$ and [10] does not depend on ζ. This is the result of analyzing error in terms of transportation norm, which allows coins with equivalent biases ($\zeta = 0$) to be merged and improbable coins ($\pi_{\min} \approx 0$) to be ignored without severe consequences.

The Matrix Pencil Method used by [3] gives the desired parameters of the model (θ) directly, yielding a straightforward stability analysis. While runtime is not discussed in [3], the method requires solving a generalized eigenvalue problem, which in practice requires time $O(k^3)$.[2] We propose improving runtime by instead using Prony's Method, a close "relative" of the Matrix Pencil Method. Prony's Method gives coefficients of a polynomial whose

[1] Of course, collecting samples also takes time, so one might ask whether the sample complexity should be included within the runtime term. However, the process of sampling and computing the moments or frequencies is computationally trivial. It can often be done under a very restrictive computational model, such as streaming. Or samples might be collected in parallel. Or the frequencies might be otherwise available from an external source. Thus, we make a clear distinction between two resources which a source identification algorithm requires: the sample size, which as noted, translates directly into the accuracy of the moments; and the runtime, given those moments.

[2] It is possible that the runtime could be improved to the time it takes to multiply two $k \times k$ matrices. This is still much worse than $O(k^2)$, and the best guarantees hide impractical constants.

roots lie at the desired parameters (θ). This significantly complicates the stability analysis, which becomes the main undertaking of this paper.

Our result. Our main result is to provide source identification of k-coin mixtures using the Prony method, simultaneously requiring moment accuracy only $\pi_{\min}^2 \zeta^{O(k)}$, and achieving runtime $O(k^{2+o(1)})$.

The main technical contributions needed for the result are Theorem 16 and Lemma 28, which give quantitative characterizations of the error propagation occurring in Prony's method. Recall that Prony's method interprets the kernel of a Hankel matrix as a polynomial with roots corresponding to the support of $\mathcal{P}$. We are able to show that the approximate Hankel matrix has a pseudo-kernel that is close to the kernel of the exact Hankel matrix; this pseudo-kernel has roots close to the roots of the exact kernel when interpreted as a polynomial.

Our result also implies an improvement in identifying pure topic models, via the reductions in [9], [10]. These reductions require solving k binary instances, and the required accuracy of those solutions necessitates a post-reduction moment accuracy of $\exp(-\Omega(k^2 \log k))$ in both papers (hence, required sample size $\exp(O(k^2 \log k))$). Our result improves the required accuracy the moments to $\exp(-\Omega(k \log k))$ (sample size $\exp(O(k \log k))$) and the runtime to $O(k^{3+o(1)})$. A more detailed comparison with previous work on topic models is given in Section 2.6.

We have posted a working implementation of the algorithm on the following public Jupyter Notebook: Online notebook implementation.[3] (Tested in Chrome and Safari.)

Related work. The k-coin problem becomes easier when m is superlinear in k, and trivial when m is $\Omega(k^2 \log k)$. Therefore, we focus on the smallest m for which the problem is solvable, which is $m = 2k - 1$ if k is assumed, or $m = 2k$ if k needs to be verified. As noted previously, three prior papers gave algorithms with worse performance than ours. Roughly stating the results (ignoring dependence on ζ and on $\pi_{\min}$), they are as follows. The paper [9] solved the problem with moment accuracy $k^{-O(k^2)}$ (sample size

[3] https://colab.research.google.com/drive/1qR6VOYSjqO8LPxqHhyY0ap_VL1apt9yS?usp=sharing

$k^{O(k^2)}$) and runtime $\mathrm{poly}(k)$. That paper also proves a lower bound of $\exp(k)$ on the sample size needed to solve the problem. Subsequently, a different solution using near optimal sample size $s = k^{O(k)}$, but much worse runtime of $k^{O(k^2)}$, was given in [10]. More recently, an algorithm achieving sample size $s = k^{O(k)}$ and runtime of $\mathrm{poly}(k)$ was analyzed in [3]. (Incidentally all of these papers use $m = 2k - 1$, and hence do not deal with verifying that the source is a k-coin distribution.)

In [9], [10], the k-coin problem arises as the output of a reduction from the problem of identifying topic models, introduced in [11], [12]. A (pure) k-topic model is simply analogous to the k-coin problem with highly multi-sided coins. There has been ample work on learning pure and mixed topic models, under various restrictive assumptions on the model, and also without restrictions [9], [10], [13], [14]. The reductions of [9], [10] can be used in conjunction with our algorithm to reduce the sample size and runtime required to solve the topic model problem. This is discussed in Section 2.6.

In [3], the k-coin problem arises as output of a reduction from the problem of inferring population histories (see the references therein). Our results improve both the sample size and the runtime complexity of the solution. We do note that the k-coin algorithm in [3] could have been used in conjunction with the reductions in [9], [10] to solve the topic model problem; the bounds derived this way would be worse than the bounds we prove in this paper.

We also mention some generalizations of the k-coin problem that were considered in the literature. Most obvious is mixtures of k product distributions on $\{0,1\}^m$. That is, the formulation is the same as ours except that $X_1, \ldots, X_m$ are merely required to be independent, but not necessarily iid, conditional on the hidden variable H. This problem has been the focus of considerable research in the past two decades [15]–[20]. Clearly, in this case a larger m is no longer purely helpful, since the number of degrees of freedom of the problem also goes up with m. It should be noted, though, that the strongest results in this sequence, [19] and [20], do not address the problem of *identifying* the source model; rather, they *learn* a model which generates similar statistics. On the positive side, this task can sometimes be performed even under conditions where there is not enough information in the statistics for identification (i.e., when there are models with near-enough statistics that are far apart in, say, transportation distance); but on the negative side, since

these algorithms (as well as the algorithm in [10]) are forced to perform an exhaustive enumeration over a large grid of potential models, their computational efficiency does not much improve even when the statistics are known to sufficiently-good accuracy that only a very small-diameter (in transportation distance) set of models could generate them.

The distinction between the "identification" and "learning" goals was made already in [16], who solved the identification problem for mixtures of $k = 2$ product distributions on $\{0,1\}^m$. Similar results for somewhat more general models were achieved at a similar time in [17]. The best result to date [20] learns in time $k^{k^3} \cdot m^{O(k^2)}$, improving upon a previous result [19] of $m^{O(k^3)}$. The same paper [20] shows a lower bound of $m^{\Omega(\sqrt{k})}$ on the sample size of the task.

Beyond mixtures of product distributions, an even more complex but important class of source identification problems arises when the hidden variable (our "H") may be just one of several such variables, and when a known directed causal structure exists among the observed variables (the "X_i"). This is a very broad field of investigation and we point only to [5], [21] for background, and to [6] for an example of how (with some additional assumptions on the distributions of the X_i) certain models can be handled.

2.2 Mixture Models and Other Definitions

We specialize the definitions appearing in Chapter 1 to this particular problem to reduce notational clutter.

Definition 6 (The k-coin model)**.** A k-coin model $\mathcal{M} = (\theta, \pi)$ is a mixture, with non-negative mixing weights $\pi_1, \ldots, \pi_k$, of k Bernoulli variables with success probabilities $\theta_1, \ldots, \theta_k$, respectively.

Definition 7 (m-snapshots of a k-coin model)**.** Given a k-coin model $\mathcal{M} = (\theta, \pi)$, an *$m$-snapshot* is a sample from the mixture of binomial distributions $\pi_1 \,\mathrm{Binomial}(m, \theta_1) + \ldots + \pi_k \,\mathrm{Binomial}(m, \theta_k)$. (The binomial is a sufficient statistic for m rv's $X_1, \ldots, X_m$ because they are iid given the selected coin.)

For a k-coin model, the moments defined in equation (2.3) can be written as follows where δ_θ is the Dirac measure at θ,

$$\mathcal{P} = \pi_1 \delta_{\theta_1} + \cdots + \pi_k \delta_{\theta_k}, \quad \mu_i = \sum_{j=1}^{k} \theta_j^i \pi_j.$$

Definition 8 (Separation for polynomials and mixtures)**.** For a k-coin probability model $\mathcal{M} = (\theta, \pi)$, define the separation by $\zeta(\mathcal{M}) = \min_{i \neq j} |\theta_i - \theta_j|$. For a degree k polynomial with roots $\beta_1, \ldots, \beta_k \in \mathbb{C}$, define the root separation by $\min_{i \neq j} |\beta_i - \beta_j|$.

Definition 9. The rectangular Vandermonde matrix $V_\theta^{(m)} \in \mathbb{R}^{(m+1) \times k}$ associated with a vector $\theta \in \mathbb{C}^k$ is given by

$$V_\theta^{(m)} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ \theta_1 & \theta_2 & \theta_3 & \cdots & \theta_k \\ \theta_1^2 & \theta_2^2 & \theta_3^2 & \cdots & \theta_k^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_1^m & \theta_2^m & \theta_3^m & \cdots & \theta_k^m \end{bmatrix}.$$

We will denote the square Vandermonde matrix by $V_\theta \coloneqq V_\theta^{(k-1)}$.

Definition 10 (Hankel Matrix)**.** The $(k+1) \times (k+1)$ Hankel matrix $\mathcal{H}_{k+1} = \mathcal{H}_{k+1}(\mathcal{P})$ is defined by $\mathcal{H}_{k+1} = [\mu_{i+j}]_{i,j=0}^{k}$.

Note that if $\mathcal{P}$ is supported on a set of cardinality k (a.k.a. a k-coin distribution), then

$$\mathcal{H}_{k+1} = \sum_{j=1}^{k} \pi_j \theta_j \theta_j^{\mathsf{T}} = V_\theta^{(k)} \operatorname{diag}(\pi_1, \ldots, \pi_k) {V_\theta^{(k)}}^{\mathsf{T}} \tag{2.4}$$

where $\theta_j^{\mathsf{T}} = (1, \theta_j, \theta_j^2, \theta_j^3, \ldots, \theta_j^k)$. This also shows that the Hankel matrix is positive semi-definite.

Definition 11 (Polynomial associated with a vector)**.** We associate to each vector $q \in \mathbb{R}^k$ a degree $k-1$ polynomial $\hat{q}(x) = \sum_{j=0}^{k-1} q_j x^j$. (For this reason we use zero indexing for the vector.)

Definition 12. For a matrix M, let $\|M\|_2$ denote the $2 \to 2$ operator norm of M. Thus, $\|M\|_2 = \sigma_{\max}(M)$, the largest singular value of M.

Definition 13. For a Hermitian matrix M, let $\lambda_i(M)$ denote the ith smallest eigenvalue of M. In particular $\lambda_1(M)$ is the smallest eigenvalue of M.

Definition 14 (Euclidean projection onto a closed convex set)**.** For a closed convex set $S \subseteq \mathbb{R}^k$ and any point $x \notin S$, the Euclidean projection of x onto S is $\operatorname{Proj}_S(x) \coloneqq \arg\min_{y \in S} \|y - x\|_2$. This projection is unique.

2.3 Properties of Hankel Matrices

We begin with some properties of Hankel matrices corresponding to finitely supported distributions that follow from results in Chihara [22]. (See Schmüdgen [23, Ch. 10] for a complete characterization.) For completeness, a proof is provided in Section 2.9.

Lemma 15. *Let $\mathcal{P}$ be a probability measure on $[0,1]$. Then,*

1. *$\mathcal{P}$ is supported on a set of cardinality at most k iff $\mathcal{H}_{k+1}$ is singular.*
2. *If the support of $\mathcal{P}$ is a set $\{\theta_1, \ldots, \theta_k\} \subset [0,1]$ then the kernel of $\mathcal{H}_{k+1}$ is spanned by the vector $q \in \mathbb{R}^{k+1}$ where $\hat{q}(z) = \prod_{i=1}^{k}(z - \theta_i)$ is the unique monic polynomial with roots at the support of $\mathcal{P}$.*

The above lemma is classic in the theory of Orthogonal Polynomials, but for completeness we provide a proof in Section 2.9. An essential contribution of this chapter is to strengthen the above lemma with a *quantitative* version:

Theorem 16. *Let $\mathcal{P} = (\theta, \pi)$ be a k-coin distribution with separation ζ, and let $\mathcal{H}_k := \mathcal{H}_k(\mathcal{P})$. Then, with I representing the $k \times k$ identity matrix and $\succeq$ being the PSD order,*

$$\mathcal{H}_k \succeq \frac{\pi_{\min}}{k} \cdot \left(\frac{\zeta}{8}\right)^{2k-2} \cdot I \ .$$

Proof. We must show that for every monic degree $k' \leq k-1$ polynomial $\hat{q}$, represented by $q \in \mathbb{R}^k$,

$$q^{\mathsf{T}} \mathcal{H}_k q \geq \frac{\pi_{\min}}{k} \cdot \left(\frac{\zeta}{8}\right)^{2k-2} \cdot \|q\|_2^2 .$$

Let $\beta_1, \beta_2, \ldots, \beta_{k'}$ be the roots (possibly complex) of the polynomial $\hat{q}$, ordered so that $|\beta_1| \geq |\beta_2| \geq \cdots \geq |\beta_{k'}|$. Since $\hat{q}$ is monic, we can write $\hat{q}(x) = \prod_{j=1}^{k'}(x - \beta_j)$. As the balls $B(\theta_i, \zeta/2)$, $i = 1, 2, \ldots, k$, are disjoint, by the pigeonhole principle, there exists an $i \in \{1, 2, \ldots, k\}$ such that $B(\theta_i, \zeta/2) \cap \{\beta_1, \beta_2, \ldots, \beta_{k'}\} = \emptyset$. The value of $\hat{q}$ at θ_i is

$$\hat{q}(\theta_i) = \prod_{j=1}^{k'}(\theta_i - \beta_j).$$

There must be some $\ell \in \{0, 1, 2, \ldots, k'\}$ such that $|q_\ell|^2 \geq \frac{\|q\|_2^2}{k'+1}$. Notice that $|q_\ell| = |e_{k'-\ell}(\beta_1, \beta_2, \ldots, \beta_{k'})|$, where e_r is the r-th elementary symmetric polynomial over k' variables. ($e_0 = 1, e_1 = \sum \beta_i, e_2 = \sum_{i<j} \beta_i \beta_j$, etc...) So, $e_{k'-\ell}$ is the sum over $\binom{k'}{k'-\ell} \leq 2^{k'}$ monomials, hence $|\beta_1 \beta_2 \cdots \beta_{k'-\ell}| \geq \frac{\|q\|_2}{(\sqrt{k'+1})2^{k'}}$. Eliminating from the product all the factors whose absolute value is below 2, we get that for some $r \leq k' - \ell$, $|\beta_1 \beta_2 \cdots \beta_r| \geq \frac{\|q\|_2}{(\sqrt{k'+1})4^{k'}}$. For $j \in \{1, 2, \ldots, r\}$, since $|\beta_j| \geq 2$ and $\theta_i \in [0, 1]$, it follows that $|\theta_i - \beta_j| \geq \frac{|\beta_j|}{2}$. Also, by the definition of i we have that $|\theta_i - \beta_j| > \zeta/2$ for all $j \in \{1, 2, \ldots, k'\}$. Thus, we have that

$$\begin{aligned} |\hat{q}(\theta_i)| &= \left(\prod_{j=1}^{r} |\theta_i - \beta_j|\right)\left(\prod_{j=r+1}^{k'} |\theta_i - \beta_j|\right) \geq \frac{|\beta_1 \beta_2 \cdots \beta_r|}{2^r}\left(\frac{\zeta}{2}\right)^{k'-r} \\ &\geq \frac{\|q\|_2}{(\sqrt{k'+1})8^{k'}}\zeta^{k'} \geq \frac{1}{\sqrt{k}} \cdot \left(\frac{\zeta}{8}\right)^{k-1} \|q\|_2 . \end{aligned}$$

Therefore,

$$q^\mathsf{T} \mathcal{H}_k q = \sum_{j=1}^{k} \pi_j \cdot (\hat{q}(\theta_j))^2 \geq \pi_{\min} \cdot (\hat{q}(\theta_i))^2 > \pi_{\min} \cdot \frac{1}{k} \cdot \left(\frac{\zeta}{8}\right)^{2k-2} \cdot \|q\|_2^2 .$$

□

Corollary 17. *For a k-coin model (θ, π), $\lambda_2(\mathcal{H}_{k+1}) > \pi_{\min} \cdot \left(\frac{\zeta}{16}\right)^{2k-2}$.*

Proof. By the Courant-Fischer-Weyl min-max principle, the smallest eigenvalue of $\mathcal{H}_k$ is given by minimizing the Rayleigh-Ritz quotient. Let $q \neq 0$ be a minimizer of $\frac{q^\mathsf{T} \mathcal{H}_k q}{q^\mathsf{T} q}$. Let k' be greatest such that $q_{k'} \neq 0$, and w.l.o.g. set $q_{k'} = 1$. Then by Theorem 16,

$$\lambda_1(\mathcal{H}_k) = \min_{q \neq 0} \frac{q^\mathsf{T} \mathcal{H}_k q}{q^\mathsf{T} q} \geq \frac{\pi_{\min}}{k} \cdot \left(\frac{\zeta}{8}\right)^{2k-2} \geq \pi_{\min} \cdot \left(\frac{\zeta}{16}\right)^{2k-2},$$

where the last inequality follows from observing that $1/k \geq 1/2^{2k-1}$ for $k \geq 2$. Notice that $\mathcal{H}_k$ is a principal submatrix of $\mathcal{H}_{k+1}$. Therefore, by the Cauchy interlacing theorem (Theorem 40), $\lambda_2(\mathcal{H}_{k+1}) \geq \lambda_1(\mathcal{H}_k)$. □

2.4 The Empirical Moments

We bound the sampling error as follows. Sample s coins and let each of the random variables h_j, $0 \leq j \leq 2k$, be the fraction of coins which came up "heads" exactly j times. Then by the additive deviation bound known as Hoeffding's inequality [24], $\Pr(|h_j - E(h_j)| \geq t) \leq 2\exp(-2t^2 s)$. Thus

Lemma 18. *If we use* $s > \frac{1}{2t^2}\log(4k/\delta)$ *samples then with probability at least* $1-\delta$*:* $\forall j, |h_j - E(h_j)| < t$.

We can convert between the normalized histogram h and the standard moments of the distribution by using the observation (Lemma 1 in [25]) that for any $t \in \mathbb{R}$,

$$t^i = \sum_{j=i}^{n} \frac{\binom{j}{i}}{\binom{n}{i}} \times \binom{n}{j} t^j (1-t)^{n-j}.$$

This gives us a linear transformation for converting from h to the vector $\tilde{\mu} = (\tilde{\mu}_0, \ldots, \tilde{\mu}_{2k})$. Define $\mathrm{Pas} \in \mathbb{R}^{(2k+1)\times(2k+1)}$ (using zero-indexing) by

$$\mathrm{Pas}_{ij} = \begin{cases} \frac{\binom{j}{i}}{\binom{2k}{i}} & \text{if } j \geq i \\ 0 & \text{otherwise;} \end{cases}$$

then $\tilde{\mu} = \mathrm{Pas}\, h$.

Lemma 19. $\|\mathrm{Pas}\|_2 \leq 6^k$. Proof in Appendix 2.9.

Now let $\mu = (\mu_0, \ldots, \mu_{2k})$ be the actual vector of moments of the distribution $\mathcal{P}$.

Lemma 20. *For every* $\varepsilon > 0$*, using* $s = 2^{O(k)} \cdot \frac{1}{\varepsilon^2} \cdot \log(1/\delta)$ *samples gives us estimated moments* $\tilde{\mu} = (\tilde{\mu}_0, \ldots, \tilde{\mu}_{2k})$ *satisfying* $\|\tilde{\mu} - \mu\|_\infty \leq \varepsilon$ *with probability at least* $1-\delta$.

Proof. Follows directly from Lemma 19 and Lemma 18. □

Given an s-sample as above with empirical moments $\tilde{\mu}_0, \tilde{\mu}_1, \ldots, \tilde{\mu}_{2k}$, denote by $\tilde{\mathcal{H}}_{k+1}$ the empirical Hankel matrix,

$$\tilde{\mathcal{H}}_{k+1} = \begin{bmatrix} \tilde{\mu}_0 & \tilde{\mu}_1 & \tilde{\mu}_2 & \cdots & \tilde{\mu}_k \\ \tilde{\mu}_1 & \tilde{\mu}_2 & \tilde{\mu}_3 & \cdots & \tilde{\mu}_{k+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tilde{\mu}_k & \tilde{\mu}_{k+1} & \tilde{\mu}_{k+2} & \cdots & \tilde{\mu}_{2k} \end{bmatrix}. \tag{2.5}$$

Corollary 21. *For every* $\varepsilon > 0$*, using* $s = 2^{O(k)} \cdot \frac{1}{\varepsilon^2} \cdot \log(1/\delta)$ *samples, we can obtain an empirical Hankel matrix satisfying* $\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_2 \leq \varepsilon$ *with probability at least* $1-\delta$.

Proof. We have $\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_2 \leq \left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_{\mathrm{F}} \leq (k+1) \cdot \|\tilde{\mu} - \mu\|_\infty$. Now use Lemma 20 with $\frac{\varepsilon}{k+1}$. □

2.5 Learning the Source

In this section, we define our learning algorithm, and we state and prove our main result and applications. The auxiliary lemmas are stated and proved in Section 2.7. The algorithm is specified given k, lower bounds on the source parameters ζ and $\pi_{\min}$, the empirical histogram h, and a parameter γ controlling the output accuracy. The algorithm is a straightforward implementation of Prony's method with numerical approximations where needed. See Algorithm 1 for the full description of the algorithm (where the parameter for probability of success, $1-\delta$, has been suppressed in favor of a constant "0.99"). On line 5, we take the output of Prony's method and project the roots back into $[0,1]$ and on line 6 in RectifyWeights we do a similar correction to ensure that the weights are non-negative and still sum to one. While simple, we defer the implementation of RectifyWeights to Algorithm 2 in Appendix 2.8.

Procedure `LearnPowerDistribution`$(k, \zeta, \pi_{\min}, \tilde{\mathcal{H}}_{k+1}, \gamma)$**:**

1 $v \xleftarrow{\varepsilon_1-\text{approx}} \arg\min\{v^{\mathsf{T}}\tilde{\mathcal{H}}_{k+1}v : \ v^{\mathsf{T}}v = 1\}$
// $\varepsilon_1 = \pi_{\min} \cdot 2^{-\gamma} \cdot (\zeta/16)^{2k}$

2

3 $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_k \xleftarrow{\varepsilon_2-\text{approx}} \text{roots}(\hat{v})$
// $\varepsilon_2 = \frac{1}{6k} \cdot 2^{-\gamma} \cdot (\zeta/2)^k$

4

5 $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k \leftarrow \text{Proj}_{[0,1]}(\tilde{\beta}_1), \ldots, \text{Proj}_{[0,1]}(\tilde{\beta}_k)$

6 $\tilde{\pi} \leftarrow \text{RectifyWeights}(V_{\tilde{\theta}}^{-1}\tilde{\mu})$
`// see Algorithm 2 on page 24`

7

8 **return** $\tilde{\mathcal{M}} \leftarrow (\tilde{\theta}, \tilde{\pi})$

Procedure `LearnCoinMixture`$(k, \zeta, \pi_{\min}, \gamma, h)$**:**

9 $\tilde{\mu} \leftarrow \text{Pas}\, h$

10 $\tilde{\mathcal{H}}_{k+1} \leftarrow \text{Hankel}(\tilde{\mu})$

11 **return** $\tilde{\mathcal{M}} \leftarrow$ `LearnPowerDistribution`$(k, \zeta, \pi_{\min}, \tilde{\mathcal{H}}_{k+1}, \gamma)$

Algorithm 1: The main learning algorithm.

Theorem 22. *Let $\mathcal{M} = (\theta, \pi)$ be a k-coin model with separation $\zeta = \zeta(\mathcal{M})$. Given an empirical Hankel matrix $\tilde{\mathcal{H}}_{k+1}$ satisfying $\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_2 \leq \pi_{\min} \cdot 2^{-\gamma} \cdot (\zeta/16)^{4k}$, the procedure LearnPowerDistribution in Algorithm 1 outputs a model $\tilde{\mathcal{M}} = (\tilde{\theta}, \tilde{\pi})$ satisfying*

$$\left\|\theta - \tilde{\theta}\right\|_\infty, \|\pi - \tilde{\pi}\|_\infty \leq 2^{-\gamma}$$

using $O(k^2 \log k + k \log^2 k \cdot \log(\log \zeta^{-1} + \log \pi_{\min}^{-1} + \gamma))$ *arithmetic operations.*

Proof. Throughout the proof, we make no attempt to optimize the absolute constants that are used. Let u_1 denote the unit vector spanning the kernel of $\mathcal{H}_{k+1}$, and let v_1 denote the eigenvector corresponding to the smallest eigenvalue of $\tilde{\mathcal{H}}_{k+1}$. Also, let $\varepsilon_0 > 0$ be a sufficiently small constant, to be determined later. The analysis of LearnPowerDistribution can be broken down into steps, each of which degrades the initial accuracy obtained for the Hankel matrix. The outline is as follows. The auxiliary claims and proofs appear mostly in Section 2.7.

1. As $\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_2 \leq \pi_{\min} \cdot 2^{-\gamma} \cdot (\zeta/16)^{4k}$, by Lemma 25,
$$\|u_1 - v_1\|_2 < \sqrt{2(k+1)} \cdot 2^{-\gamma} \cdot (\zeta/16)^{2k} < \frac{1}{2} \cdot 2^{-\gamma} \cdot (\zeta/8)^{2k}.$$

2. We use Lemma 26 with $\varepsilon = \pi_{\min} \cdot 2^{-\gamma} \cdot (\zeta/16)^{2k}$, which satisfies the conditions of the lemma. We compute $v \in \mathbb{R}^{k+1}$ such that
$$\|v - v_1\|_2 \leq \varepsilon < \frac{1}{2} \cdot 2^{-\gamma} \cdot (\zeta/16)^{2k},$$
using $O(k^2 \log k + k \log^2 k \cdot \log(\log \zeta^{-1} + \log \pi_{\min}^{-1} + \gamma))$ arithmetic operations.

3. As $\|u_1 - v_1\|_2, \|v - v_1\|_2 < \frac{1}{2} \cdot 2^{-\gamma} \cdot (\zeta/16)^{2k}$, we have that $\|u_1 - v\|_2 < 2^{-\gamma} \cdot (\zeta/16)^{2k}$. So, by Lemma 27,
$$\|q - r\|_\infty < 2^k \cdot \sqrt{k+1} \cdot 2^{-\gamma} \cdot (\zeta/16)^{2k} < 2^{-\gamma} \cdot (\zeta/8)^{2k},$$
where $q := u_1 / |(u_1)_k|$, $r := v / |(u_1)_k|$.

4. As $\|q - r\|_\infty < 2^{-\gamma} \cdot (\zeta/8)^{2k} \leq \frac{1}{24k(k+1)} \cdot 2^{-\gamma} \cdot (\zeta/2)^{2k-1}$, by Lemma 28 we have that
$$d(\theta, \beta) \leq \frac{1}{6(k+1)} \cdot 2^{-\gamma} \cdot (\zeta/2)^k$$
(where θ is the vector of roots of $\hat{q}$ and β is the vector of roots of $\hat{r}$ and d is the matching distance, defined in Lemma 28).

5. We use Corollary 43 with $\rho = \varepsilon = \frac{1}{6(k+1)} \cdot 2^{-\gamma} \cdot (\zeta/2)^k$, which satisfy the conditions of the corollary. Thus, we can compute biases $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ satisfying
$$\left\|\tilde{\theta} - \theta\right\|_\infty \leq \rho + \varepsilon \leq \frac{1}{3(k+1)} \cdot 2^{-\gamma} \cdot (\zeta/2)^k,$$

using $O(k \log^2 k \cdot (\log(\log \zeta^{-1} + \gamma) + \log^2 k))$ arithmetic operations.

6. Finally, line 6 can be executed in the time it takes to invert the Vandermonde matrix $V_{\tilde{\theta}}$ (i.e., $O(k^2)$ arithmetic operations, for instance using Parker's Algorithm [26]; by Lemma 34, the procedure RectifyWeights takes $O(k)$ operations). By Corollary 35, as $\|\tilde{\theta} - \theta\|_\infty, \|\tilde{\mu} - \mu\|_\infty \leq \frac{1}{3(k+1)} \cdot 2^{-\gamma} \cdot (\zeta/2)^k$ (the guarantee for $\tilde{\mu}$ is implied with plenty of room to spare by our assumption on the sample), we have $\|\tilde{\pi} - \pi\|_\infty \leq 2^{-\gamma}$.

□

Corollary 23. *Let $\mathcal{M} = (\theta, \pi)$ be a k-coin model with separation $\zeta = \zeta(\mathcal{M})$. For any $\gamma \geq 1$, the procedure LearnCoinMixture in Algorithm 1 uses a histogram h for a sample of $2k$-snapshots of size $s = \pi_{\min}^{-2} \cdot 2^{O(k+\gamma)} \cdot \zeta^{-O(k)} \cdot \log \delta^{-1}$, and outputs a model $\tilde{\mathcal{M}} = (\tilde{\theta}, \tilde{\pi})$ satisfying*

$$\left\|\theta - \tilde{\theta}\right\|_\infty, \|\pi - \tilde{\pi}\|_\infty \leq 2^{-\gamma}$$

with probability at least $1 - \delta$. After sampling, LearnCoinMixture computes the approximate model $\tilde{\mathcal{M}}$ using $O(k^2 \log k + k \log^2 k \cdot \log(\log \zeta^{-1} + \log \pi_{\min}^{-1} + \gamma))$ arithmetic operations.

Proof. By Lemma 20 and Corollary 21 we can achieve $\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_2 \leq \pi_{\min} \cdot 2^{-\gamma} \cdot (\zeta/16)^{4k}$ for a sample of size $s = \pi_{\min}^{-2} \cdot 2^{O(k+\gamma)} \cdot \zeta^{-O(k)} \cdot \log \delta^{-1}$, with probability at least $1 - \delta$. The theorem immediately follows from Theorem 22. □

Notice that the proof actually gives a stronger guarantee for $\left\|\tilde{\theta} - \theta\right\|_\infty$, which is relative to $(\zeta/2)^k$. We can get a relative guarantee $\|\tilde{\pi} - \pi\|_\infty \leq \pi_{\min} \cdot 2^{-\gamma}$ by increasing the sample size by a factor of $\pi_{\min}^{-2}$.

Corollary 24. *Let $W(\mathcal{M}, \tilde{\mathcal{M}})$ denote the Wasserstein distance between models $\mathcal{M}$ and $\tilde{\mathcal{M}}$ (viewed as metric measure spaces on $[0, 1]$). Then, $W(\mathcal{M}, \tilde{\mathcal{M}}) \leq (k+1) \cdot 2^{-\gamma}$ with probability at least 0.99.* Proof in Appendix 2.9.

2.6 Implications for Topic Models

Corollary 23 improves upon the upper bound of Theorem 5.1 in [9], which uses a sample of $(2k-1)$-snapshots of size $\max\left\{(2/\zeta)^{O(k)}, (2^\gamma k)^{O(k^2)}\right\}$ to achieve accuracy $2^{-\gamma}$ with high probability, using runtime of $O(k^c)$ arithmetic

operations, for a relatively large constant c (in particular, the algorithm solves a convex quadratic program whose representation uses k^3 bits). Corollary 23 also improves upon the upper bound of [10].[4] That algorithm uses a sample size comparable to ours, but requires runtime $(2^\gamma k)^{O(k^2)}$ to achieve accuracy $2^{-\gamma}$ with high probability.

These improvements imply immediately a similar improvement for learning pure k-topic models, using known reductions from k-topic models to k-coin models. The reduction in Theorem 4.1 of [9] uses a sample of 1- and 2-snapshots of size $O\left(n \cdot \mathrm{poly}\left(\log n, k, \pi_{\min}^{-1}, \zeta^{-1}, 2^\gamma\right)\right)$, and runtime polynomial in the sample size, to reduce the problem to solving k instances of the k-coin problem with accuracy $\min\left\{(2^\gamma k/\pi_{\min}\zeta)^{-O(1)}, (2^\gamma k)^{-O(k)}\right\}$. The reduction in [10] [5] uses a sample of 1- and 2-snapshots of size $\mathrm{poly}\,(n, k, 2^\gamma)$, and runtime polynomial in the sample size, to reduce the problem to solving at most k instances of the k-coin problem with accuracy $(2^\gamma k)^{-O(k)}$. Notice that solving the k-coin outcome of either one of the two reductions using either one of the two previous algorithms requires a sample size of at least $k^{O(k^2)}$ (on account of the required accuracy). Our algorithm enables a solution to the outcome of these reductions using a sample size of $k^{O(k)}$ (and total runtime of $O(k^{3+o(1)})$). We note that the accuracy in [9], [10] is stated in terms of Wasserstein distance, which is a weaker guarantee than the one we use here (see Corollary 24).

2.7 Analysis

In this section we prove the lemmas that are needed in the proof of Theorem 22 and Corollary 23. We have to cope with the fact that roots of polynomials (and even, generally, of polynomials with well-separated roots), are notoriously ill-conditioned in terms of the polynomial coefficients [27]. For this reason we will be developing bounds specifically adapated to our situation. We begin with an estimate on the accuracy of the recovered kernel of the Hankel matrix.

Lemma 25. *Let $\mathcal{P}$ be any k-coin distribution with separation ζ. Then, for every $\varepsilon < \pi_{\min} \cdot \left(\frac{\zeta}{16}\right)^{2k}$ the following holds. Suppose that $\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_2 \leq \varepsilon$. Let u_1 be the unit vector in the kernel of $\mathcal{H}_{k+1}$ and let v_1 be the unit eigenvector corresponding to*

[4] See Theorem 3.9 in the ArXiv version

[5] See Theorem 6.1 in the ArXiv version.

$\lambda_1(\tilde{\mathcal{H}}_{k+1})$ *(chosen so that* $u_1^\mathsf{T} v_1 \geq 0$*). Then* $\|u_1 - v_1\|_2 < \sqrt{2(k+1)} \cdot \left(\frac{16}{\zeta}\right)^{2k} \cdot \frac{\varepsilon}{\pi_{\min}}$. Proof in Appendix 2.9.

Recall that $\left(\lambda_1(\tilde{\mathcal{H}}_{k+1}), v_1\right)$ is an eigenpair of $\tilde{\mathcal{H}}_{k+1}$. We need to compute a good approximation of v_1. This can be done using the following lemma. The result is implied by the algorithm of Pan and Chen (Theorem 1.2 of [28]). Extracting our lemma from the result in that paper is somewhat involved and we provide in Appendix 2.9 a brief outline of the argument (in particular, the parts that are not spelled out in that paper).

Lemma 26. *For every* ε *such that* $0 < \varepsilon < \frac{1}{2}\min\{\lambda_2(\tilde{\mathcal{H}}_{k+1}) - \lambda_1(\tilde{\mathcal{H}}_{k+1}), 1\}$*, we can compute a unit vector* v *satisfying* $\|v - v_1\|_2 < \varepsilon$ *using* $O\left(k^2 \log k + k \log^2 k \log\log(1/\varepsilon)\right)$ *arithmetic operations.*

We need to show that our computed first eigenvector of the empirical Hankel matrix is close to the kernel eigenvector of the true Hankel matrix.

Lemma 27. *Let* $\mathcal{P}$ *be any* k*-coin distribution with separation* ζ*. Let* u_1 *be a unit vector in the kernel of* $\mathcal{H}_{k+1}$*. Let* v *be a unit vector satisfying* $\|u_1 - v\|_2 < \varepsilon$ *for some* $\varepsilon > 0$*. Let* $q = u_1 / |(u_1)_k|$ *and let* $r = v / |(u_1)_k|$*. Then* $\|q - r\|_\infty < 2^k \sqrt{k+1} \cdot \varepsilon$. Proof in Appendix 2.9.

We are going to use the roots of the polynomial $\hat{r}$ as our guessed coin biases (after projecting the roots back to $[0, 1]$). We first need to show that the roots of $\hat{q}$ are well-behaved with respect to perturbations of q so that when q and r are close the roots of $\hat{q}$ are close to the roots of $\hat{r}$.

Lemma 28. *Let* $q \in \mathbb{R}^{k+1}$ *be the vector representing a degree-*k *monic polynomial with roots* $\theta_1, \theta_2, \ldots, \theta_k$ *contained in* $[0, 1]$*. Let* ζ *be the root separation for* $\hat{q}$*. Let* $r \in \mathbb{R}^{k+1}$ *represent another degree-*k *polynomial. Let* $\varepsilon \in \left(0, \frac{(\zeta/2)^k}{4k}\right)$*. If* r *satisfies* $\|q - r\|_\infty \leq \varepsilon$*, then the (possibly complex) roots* $\beta_1, \beta_2, \ldots, \beta_k$ *of* $\hat{r}$ *satisfy* $d(\theta, \beta) \leq \frac{4k\varepsilon}{(\zeta/2)^{k-1}}$ *where* $d(\theta, \beta)$ *is the optimal matching distance defined by* $d(\theta, \beta) \coloneqq \min_{\sigma \in \mathbb{S}_k} \max_i \left|\theta_i - \beta_{\sigma(i)}\right|$.

Proof. Fix any root θ_i of $\hat{q}$, and consider the ball

$$B_i = B\left(\theta_i, \frac{4k\varepsilon}{(\zeta/2)^{k-1}}\right)$$

in the complex plane. By assumption, $\frac{4k\varepsilon}{(\zeta/2)^{k-1}} < \frac{\zeta}{2}$, so there are no other roots of $\hat{q}$, aside from θ_i, in B_i. Moreover, for any $x \in B_i$, and for any $j \neq i$, we have that $|x - \theta_j| \geq \frac{\zeta}{2}$. Thus for every $x \in \partial B_i$, we have

$$|\hat{q}(x)| = \left|(x - \theta_i)\prod_{j \neq i}(x - \theta_j)\right| > \frac{4k\varepsilon}{(\zeta/2)^{k-1}}\left(\frac{\zeta}{2}\right)^{k-1} = 4k\varepsilon.$$

On the other hand, we also have that $B_i \subset B(0, (2k-1)/(2k-2))$, as $\theta_1, \ldots, \theta_k \in [0, 1]$ and $\zeta \leq \frac{1}{k-1}$. Therefore, $|x| \leq \frac{2k-1}{2k-2}$, and thus

$$|\hat{q}(x) - \hat{r}(x)| = \left|\sum_{j=0}^{k}(q_j - r_j)x^j\right| \leq \sum_{j=0}^{k}|q_j - r_j| \cdot |x|^j \leq (k+1) \cdot \left(\frac{2k-1}{2k-2}\right)^k \cdot \|q - r\|_\infty \leq 4k\varepsilon.$$

By Rouché's theorem (Theorem 41), we conclude that there is exactly one zero of $\hat{r}$ in B_i and the matching distance bound follows immediately. □

Our reconstructed coin biases will be denoted $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k$. We compute these biases by finding the roots of $\hat{v}$ (approximately), and then by projecting these roots onto the unit interval. To find the approximate roots we can use Corollary 43 in Appendix 2.10.

Once we have recovered the parameters $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$, we need to use those to recover mixture weights. This sequence of steps—first solving (approximately) for the roots, then for the mixture weights—is the essence of Prony's method [7], [1] §9.4, [8].

In Appendix 2.8, we will show that this recovery can be done by solving a linear system without paying too great a price in terms of accuracy.

2.8 Computing the Weights

Recovering the Weights

We will begin by stating results characterizing the condition number of a Vandermonde matrix under perturbations that preserve the Vandermonde structure.

Lemma 29 (Operator norm bound for a Vandermonde inverse; equation 3.2 in [29])**.** *Let $\theta \in \mathbb{R}^k$ be entry-wise non-negative, and let $q(z) = \prod_{i=1}^{k}(z - \theta_i)$. Then*

$$\left\|V_\theta^{-1}\right\|_\infty = \frac{|q(-1)|}{\min_i \{(1+\theta_i)\,|q'(\theta_i)|\}}.$$

Claim 30. *For roots $\theta_1, \ldots, \theta_j$ satisfying $|\theta_i - \theta_j| \geq \zeta$, we have $\left\|V_\theta^{-1}\right\|_\infty \leq 2^k/\zeta^{k-1}$.*

Proof. We apply Lemma 29 and observe that $|q(-1)| \leq 2^k$ and $q'(\theta_i) \geq \zeta^{k-1}$. □

We define the derivative matrix of the Vandermonde matrix by interpreting each entry as the evaluation of a polynomial at a point, $[V_a]_{ij} = p_i(a_j)$, where $p_i(t) = t^i$. Then $[V_a']_{ij} = p_i'(a_j) = i a_j^{i-1}$.

We will now define the condition number of the system,

$$\mathrm{cond}_\infty(a,b) := \lim_{\varepsilon \to 0} \sup_{\substack{\|\Delta a\|_\infty \leq \varepsilon \\ \|\Delta b\|_\infty \leq \varepsilon}} \left\{ \frac{\|\Delta x\|_\infty}{\varepsilon} \;\middle|\; V(a+\Delta a)(x+\Delta x) = b + \Delta b \right\}. \tag{2.6}$$

We will utilize a bound from [30]. After instantiating the theorem with the parameters relevant to our problem, the bound is the following:

Theorem 31 (Theorem 2.2 of [30])**.** $\mathrm{cond}_\infty(a,b) \leq \|V_a^{-1}\|_\infty + \|V_a^{-1} V_a' \,\mathrm{diag}(x)\|_\infty$.

Lemma 32. *Let $\theta \in [0,1]^k$, and let $w \in \mathbb{R}^k$ be a probability distribution over $[k]$. Let $\mu = V_\theta w$. If $\zeta \leq \min_{i \neq j} |\theta_i - \theta_j|$, $\mathrm{cond}_\infty(\theta, \mu) \leq (k+1)2^k/\zeta^{k-1}$.*

Proof. We observe that

$$\begin{aligned}\left\|V_\theta^{-1}V_\theta' \operatorname{diag}(\pi)\right\|_\infty &\le \left\|V_\theta^{-1}\right\|_\infty \left\|V_\theta' \operatorname{diag}(\pi)\right\|_\infty \\ &\le 2^k/\zeta^{k-1} \left\|V_\theta' \operatorname{diag}(\pi)\right\|_\infty \\ &= 2^k/\zeta^{k-1} \max_{i\in[k-2]} (i+1) \sum_{j=1}^{k} \left|\theta_j^i \pi_j\right| \\ &\le k2^k/\zeta^{k-1}.\end{aligned}$$

Applying the bound of Theorem 31 gives the conclusion. □

Lemma 33. *Let $\theta \in [0,1]^k$ and let $w \in \mathbb{R}^k$ be a probability distribution over $[k]$. Let $\mu = V_\theta w$, and $\zeta \le \min_{i\neq j} |\theta_i - \theta_j|$. Then $\pi' := V_{\tilde\theta}^{-1}\tilde\mu$ satisfies*

$$\|\pi' - \pi\|_\infty \le \frac{(k+1)2^k}{\zeta^{k-1}} \max\left\{\left\|\tilde\theta - \theta\right\|_\infty, \|\tilde\mu - \mu\|_\infty\right\}.$$

Proof. This follows from Lemma 32 and the definition of the condition number. □

Lemma 34. *Given any weights $\pi' \in \mathbb{R}^k$ satisfying $\sum_{i=1}^k \pi_i' = 1$, the procedure* RectifyWeights*(π') outputs in time $O(k)$ a weight vector $\tilde\pi \in [0,1]^k$ satisfying the following conditions*

(i) $\sum_{i=1}^k \tilde\pi_i = 1$.

(ii) $\|\tilde\pi - \pi\|_\infty \le (k+1)\|\pi' - \pi\|_\infty$.

Corollary 35. *Letting $\tilde\pi \in [0,1]^k$ be the output of* RectifyWeights*(π') where π' is as in Lemma 33,*

$$\|\tilde\pi - \pi\|_\infty \le \frac{(k+1)^2 2^k}{\zeta^{k-1}} \max\left\{\left\|\tilde\theta - \theta\right\|_\infty, \|\tilde\mu - \mu\|_\infty\right\}.$$

Proof. Notice that the first equation in the linear system defining π' is $\sum_{i=1}^k \pi_i' = \mathbb{1}^T\pi' = \tilde\mu_0 = 1$. Thus, π' satisfies the hypothesis of Lemma 34 and the conclusion follows. □

Procedure `RectifyWeights(`π'`):`

$I^- \leftarrow \{i \mid \pi'_i < 0\}, \quad I^+ \leftarrow \{i \mid \pi'_i \geq 0\}$

$W^- \leftarrow \sum_{i \in I^-} \pi'_i, \quad W^+ \leftarrow \sum_{i \in I^+} \pi'_i$

for $i = 1, \ldots, k$ **do**

$$\tilde{\pi}_i \leftarrow \begin{cases} 0 & \text{if } i \in I^- \\ \pi'_i \left(1 + \frac{W^-}{W^+}\right) & \text{if } i \in I^+ \end{cases}$$

end

return $\tilde{\pi}$

Algorithm 2: Correct computed weights to be non-negative and sum to one.

Fixing the weights

The weights produced prior to calling RectifyWeights sum to 1 (since they satisfy the equation $1 = \mu_0 = \sum_j \tilde{\pi}_j$) but they may lie outside $[0, 1]$. To correct this, we simply make all negative weight zero and scale all non-negative weights so that the sum of the weights doesn't change.

Proof of Lemma 34. Note that in Algorithm 2, I^- denotes the indices of the negative weights, and I^+ the positive weights. W^- and W^+ denote the sums of the weights in the corresponding set of indices.

We will now analyze $\tilde{\pi}$. First, note that we maintain property (i):

$$\begin{aligned} \sum_{i=1}^{k} \tilde{\pi}_i &= \sum_{i \in I^+} \pi'_i \left(1 + \frac{W^-}{W^+}\right) \\ &= W^+ \left(1 + \frac{W^-}{W^+}\right) \\ &= W^+ + W^- = 1. \end{aligned}$$

Now we show that the weights are non-negative. Trivially, $\tilde{\pi}_i \geq 0$ for $i \in I^-$. For $i \in I^+$,

$$\begin{aligned} W^+ &= 1 - W^- \\ &= 1 + |W^-| \\ &\geq |W^-|. \end{aligned}$$

So $\pi'_i(1 + \frac{W^-}{W^+}) \geq 0$ if $i \in I^+$ as well.

We now prove (ii). We know that the true weights π lie in $[0, 1]$, so increasing the negative weights to 0 only moves them closer to their true values. Thus,

we have $|\tilde{\pi}_i - \pi_i| \leq |\pi'_i - \pi_i|$ for all $i \in I^-$. We observe that

$$\left|W^-\right| \leq \|\pi' - \pi\|_1 \leq k \|\pi' - \pi\|_\infty$$

and then that

$$|\tilde{\pi}_i - \pi'_i| = \left| \underbrace{\left(\pi'_i / W^+\right)}_{\leq 1} W^- \right| \leq k \|\pi' - \pi\|_\infty .$$

It follows that $\|\tilde{\pi} - \pi'\|_\infty \leq k \|\pi' - \pi\|_\infty$. Now we can apply the triangle inequality to get that

$$\|\tilde{\pi} - \pi\|_\infty \leq \|\tilde{\pi} - \pi'\|_\infty + \|\pi' - \pi\|_\infty \leq (k+1) \|\pi' - \pi\|_\infty .$$

To see that the runtime is $O(k)$ we observe we can compute I^- and I^+ in linear time and likewise for W^- and W^+. Each subsequent computation of $\tilde{\pi}_i$ takes constant time. □

2.9 Deferred Proofs

Proof of Lemma 15. (Part 1.) By Equation (2.4), the rank of $\mathcal{H}_{k+1}$ for a t-coin distribution is at most t, and that implies that if $t \leq k$, then $\mathcal{H}_{k+1}$ is singular. So consider a distribution $\mathcal{P}$ on $[0,1]$ that has positive mass at $k+1$ points or more. Let $q \in \mathbb{R}^{k+1}$ be a non-zero vector. We have

$$q^\mathsf{T} \mathcal{H}_{k+1} q = \int_0^1 \left(\sum_{j=0}^{k} q_j \theta^j \right)^2 \mathrm{d}\mathcal{P}(\theta) = \int_0^1 \hat{q}^2(\theta) \, \mathrm{d}\mathcal{P}(\theta).$$

There are at most k points in $[0,1]$ where the polynomial $\hat{q}$ evaluates to 0, and the total $\mathcal{P}$ measure of those points is less than 1. Thus, $q^\mathsf{T} \mathcal{H}_{k+1} q > 0$, so $\mathcal{H}_{k+1}$ is positive definite.

(Part 2.) Since $\mathcal{H}_{k+1}$ is symmetric, its kernel is spanned by q s.t. $q^\mathsf{T} \mathcal{H}_{k+1} q = 0$. In order for the above integral to evaluate to zero over $\mathcal{P}$, we need that $\hat{q}^2(\theta) = 0$ for each point $\theta \in \mathrm{supp}(\mathcal{P})$. As $\hat{q}$ is of degree $\leq k$, it is necessarily a scalar multiple of $\prod_{i=1}^{k} (z - \theta_i)$. □

Proof of Lemma 19. We first observe that

$$\mathrm{Pas} = \begin{bmatrix} \binom{0}{0}\binom{2k}{0}^{-1} & \binom{1}{0}\binom{2k}{0}^{-1} & \cdots & \binom{2k-1}{0}\binom{2k}{0}^{-1} & \binom{2k}{0}\binom{2k}{0}^{-1} \\ 0 & \binom{1}{1}\binom{2k}{1}^{-1} & \cdots & \binom{2k-1}{1}\binom{2k}{1}^{-1} & \binom{2k}{1}\binom{2k}{1}^{-1} \\ 0 & 0 & \cdots & \binom{2k-1}{2}\binom{2k}{2}^{-1} & \binom{2k}{2}\binom{2k}{2}^{-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \binom{2k}{2k}\binom{2k}{2k}^{-1} \end{bmatrix}$$

which can be factored to obtain

$$\mathrm{Pas} = \begin{bmatrix} \binom{2k}{0}^{-1} & 0 & 0 & \cdots & 0 \\ 0 & \binom{2k}{1}^{-1} & 0 & \cdots & 0 \\ 0 & 0 & \binom{2k}{2}^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \binom{2k}{2k}^{-1} \end{bmatrix} \begin{bmatrix} \binom{0}{0} & \binom{1}{0} & \cdots & \binom{2k-1}{0} & \binom{2k}{0} \\ 0 & \binom{1}{1} & \cdots & \binom{2k-1}{1} & \binom{2k}{1} \\ 0 & 0 & \cdots & \binom{2k-1}{2} & \binom{2k}{2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \binom{2k}{2k} \end{bmatrix}.$$

Now

$$\left\| \mathrm{diag}\left(\binom{2k}{0}, \binom{2k}{1}, \ldots, \binom{2k}{2k} \right)^{-1} \right\|_2 \leq 1.$$

The Frobenius norm of the latter matrix is

$$\left(\sum_{j=0}^{2k} \sum_{i=0}^{j} \binom{j}{i}^2 \right)^{1/2} = \left(\sum_{j=0}^{2k} \binom{2j}{j} \right)^{1/2} \leq \left(2k \binom{4k}{2k} \right)^{1/2} \leq (2^k 4^{2k})^{1/2} \leq 6^k$$

for $k \geq 2$. Using the sub-multiplicativity of the operator norm and the fact that the Frobenius norm upper bounds the operator norm, we get that $\|\mathrm{Pas}\| \leq 6^k$, as desired. □

Proof of Corollary 24:. Each θ_i can be matched to its corresponding $\tilde{\theta}_i$ up to weight $\min\{\pi_i, \tilde{\pi}_i\}$. The additional $|\pi_i - \tilde{\pi}_i|$ must move an additional distance of at most 1. This gives

$$\begin{aligned} W(\mathcal{M}, \tilde{\mathcal{M}}) &\leq \sum_{i=1}^{k} |\theta_i - \tilde{\theta}_i| \cdot \min\{\pi_i, \tilde{\pi}_i\} + \sum_{i=1}^{k} |\pi_i - \tilde{\pi}_i| \\ &\leq \sum_{i=1}^{k} 2^{-\gamma} \min\{\pi_i, \tilde{\pi}_i\} + \sum_{i=1}^{k} 2^{-\gamma} \\ &\leq (k+1) \cdot 2^{-\gamma}, \end{aligned}$$

using Theorem 22 and the fact that $\sum_{i=1}^{k} \min\{\pi_i, \tilde{\pi}_i\} \leq \sum_{i=1}^{k} \pi_i = 1$. □

Proof of Lemma 25. By Weyl's inequality, we have that $\lambda_1(\tilde{\mathcal{H}}_{k+1}) \leq \varepsilon$. By Corollary 17, the eigengap $\lambda_2(\mathcal{H}_{k+1}) - \lambda_1(\tilde{\mathcal{H}}_{k+1})$ is at least

$$\pi_{\min} \left(\frac{\zeta}{16} \right)^{2k-2} - \pi_{\min} \left(\frac{\zeta}{16} \right)^{2k} > \pi_{\min} \left(\frac{\zeta}{16} \right)^{2k}.$$

Now we can use Corollary 38 to obtain

$$u_1^\mathsf{T} v_1 = \left|u_1^\mathsf{T} v_1\right| \geq \left(1 - \frac{\left\|\mathcal{H}_{k+1} - \tilde{\mathcal{H}}_{k+1}\right\|_F^2}{\left|\lambda_2(\mathcal{H}_{k+1}) - \lambda_1(\tilde{\mathcal{H}}_{k+1})\right|^2}\right)^{1/2} > 1 - \frac{(k+1)\cdot\varepsilon^2}{w_{\min}^2\left(\frac{\zeta}{16}\right)^{4k}}$$
$$= 1 - (k+1)\cdot\left(\frac{16}{\zeta}\right)^{4k}\cdot\left(\frac{\varepsilon}{\pi_{\min}}\right)^2.$$

Since $\|u_1 - v_1\|_2^2 = 2 - 2u_1^\mathsf{T} v_1$ we get that

$$\|u_1 - v_1\|_2^2 < 2(k+1)\cdot\left(\frac{16}{\zeta}\right)^{4k}\cdot\left(\frac{\varepsilon}{\pi_{\min}}\right)^2.$$

□

Proof sketch of Lemma 26. We follow the outline in the papers by Pan, Chen, and Zheng [28], [31]. As $\tilde{\mathcal{H}}_{k+1}$ is a Hankel matrix, a similarity transformation $A = T\,\tilde{\mathcal{H}}_{k+1}\,T^{-1}$, where A is tridiagonal, can be computed in time $O(k^2 \log k)$. The characteristic polynomial $c_A(x)$ of A can then be computed in time $O(k)$. Then, a root $\tilde{\lambda}$ that satisfies $|\tilde{\lambda} - \lambda_1(\tilde{\mathcal{H}}_{k+1})| < \varepsilon^2$ can be computed in time $O\left((k\log^2 k)(\log\log(1/\varepsilon) + \log^2 k)\right)$ (see Theorem 42; note that $\|A\|_2 = \|\tilde{\mathcal{H}}_{k+1}\|_2$, thus it is trivially upper bounded by $(k+1)^2$). Next, proceed to compute v as follows. Pick an initial guess $v^{(0)}$ uniformly at random on the unit sphere (i.e., from the unit Haar measure on the sphere). We need $v_1^\mathsf{T} v^{(0)} > \frac{1}{\sqrt{k}}$, which happens with constant probability. To boost the success probability to $1-\delta$, we can repeat the entire process $O(\log(1/\delta))$ times. For constant δ, this does not affect the asymptotic bound. We compute $v^{(1)}, v^{(2)}, \ldots$ using the inverse power iteration (see, for instance, Chapter 4 in [32]): Solve for $\tilde{v}^{(t)}$ the system of linear equations $\left(\tilde{\lambda} I - \tilde{\mathcal{H}}_{k+1}\right)\tilde{v}^{(t)} = v^{(t-1)}$, then set $v^{(t)} = \frac{\tilde{v}^{(t)}}{\|\tilde{v}^{(t)}\|_2}$. As $\tilde{\mathcal{H}}_{k+1}$ is a Hankel matrix, this can be done using $O(k^2)$ arithmetic operations. How many iterations are needed?—It is known that if $\lambda_1(\tilde{\mathcal{H}}_{k+1})$ is the unique eigenvalue of $\tilde{\mathcal{H}}_{k+1}$ that is closest to $\tilde{\lambda}$, and if $v_1^\mathsf{T} v^{(0)} > 0$, then $\tan\theta^{(t)} \leq \rho\cdot\tan\theta^{(t-1)}$, where $\theta^{(t)}$ is the angle between v_1 and $v^{(t)}$, and $\rho = \frac{|\tilde{\lambda}-\lambda_1(\tilde{\mathcal{H}}_{k+1})|}{|\tilde{\lambda}-\lambda_2|}$, where λ_2 is an eigenvalue of $\tilde{\mathcal{H}}_{k+1}$ that is second-closest to $\tilde{\lambda}$. Notice that in our case $\rho = \frac{|\tilde{\lambda}-\lambda_1(\tilde{\mathcal{H}}_{k+1})|}{\min_{i>1}|\tilde{\lambda}-\lambda_i(\tilde{\mathcal{H}}_{k+1})|} < \frac{\varepsilon^2}{\varepsilon-\varepsilon^2} < 2\varepsilon$. As $\tan\theta^{(0)} \leq \sqrt{k}$, after $t = O(\log_{1/2\varepsilon} k)$ iterations, we have $\tan\theta^{(t)} < \varepsilon$. This implies that $\left\|v^{(t)} - v_1\right\|_2 < \varepsilon$. □

Proof of Lemma 27. Notice that q and r are well-defined, as $(u_1)_k \neq 0$ by the second part of Lemma 15. Now each of the coefficients of q can be bounded by

$$|q_i| = |e_{k-i}(\theta_1, \ldots, \theta_k)| \leq \binom{k}{i}$$

where e_r is the r-th elementary symmetric polynomial over k variables. Now $\|q\|_2 \leq \sqrt{k+1} \cdot \|q\|_1 \leq 2^k\sqrt{k+1}$. Since $|(u_1)_k| \cdot \|q\|_2 = \|u_1\|_2 = 1$, we have $|(u_1)_k| \leq \frac{1}{2^k\sqrt{k+1}}$, and

$$\|q - r\|_\infty \leq \|q - r\|_2 \leq 2^k\sqrt{k+1} \cdot \|u_1 - v\|_2 < 2^k\sqrt{k+1} \cdot \varepsilon,$$

as stipulated. □

Proof of Lemma 32. We observe that

$$\begin{aligned}
\left\|V_\theta^{-1} V_\theta' \operatorname{diag}(\pi)\right\|_\infty &\leq \left\|V_\theta^{-1}\right\|_\infty \left\|V_\theta' \operatorname{diag}(\pi)\right\|_\infty \\
&\leq 2^k/\zeta^{k-1} \left\|V_\theta' \operatorname{diag}(\pi)\right\|_\infty \\
&= 2^k/\zeta^{k-1} \max_{i \in [k-2]} (i+1) \sum_{j=1}^{k} \left|\theta_j^i \pi_j\right| \\
&\leq k2^k/\zeta^{k-1}.
\end{aligned}$$

Applying the bound of Theorem 31 gives the conclusion. □

2.10 Useful Theorems

Consider two $n \times n$ Hermitian matrices A, B, with spectral decompositions $A = \sum_{i=1}^n \kappa_i u_i u_i^\mathsf{T}$ and $B = \sum_{i=1}^n \lambda_i v_i v_i^\mathsf{T}$, where the eigenvalues of both matrices are sorted in increasing order (i.e., $\kappa_1 \leq \kappa_2 \leq \cdots \leq \kappa_n$ and $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$). Also, let $P = B - A$ and let $\rho_1 \leq \rho_2 \leq \cdots \leq \rho_n$ be the eigenvalues of P in increasing order.

Theorem 36 (Weyl's inequality)**.** *For every $i \in \{1, 2, \ldots, n\}$,*

$$\kappa_i + \rho_1 \leq \lambda_i \leq \kappa_i + \rho_n.$$

Theorem 37 (Davis-Kahan $\sin\Theta$ theorem)**.** *Using the above definitions, let i_0, i_1 be integers such that $1 \leq i_0 \leq i_1 \leq n$, and let*

$$g = \inf\{|\kappa - \lambda| : \ \kappa \in [\kappa_{i_0}, \kappa_{i_1}] \wedge \lambda \in (-\infty, \lambda_{i_0-1}] \cup [\lambda_{i_1+1}, +\infty)\},$$

where we define $\lambda_0 = -\infty$ *and* $\lambda_{n+1} = \infty$*. Then,*

$$\|\sin\Theta(U,V)\|_F \leq \frac{\|P\|_F}{g},$$

where U (V*, respectively*) *is the* $n \times i_1 - i_0 + 1$ *matrix whose columns are* $u_{i_0}, \ldots, u_{i_1}$ ($v_{i_0}, \ldots, v_{i_1}$*, respectively*)*,* $\Theta(U,V)$ *is the* $i_1 - i_0 + 1 \times i_1 - i_0 + 1$ *diagonal matrix whose* i*-th diagonal entry is the* i*-th principal angle between the column spaces of* U *and* V*, and* $\sin\Theta(U,V)$ *is the diagonal matrix derived by applying the function* sin *entrywise to* $\Theta(U,V)$*. The same inequality holds if the Frobenius norm is replaced by any orthogonally invariant norm, e.g., an operator norm* $\|\cdot\|_{\mathrm{op}}$*.*

Corollary 38. *Using the same definitions,*

$$|u_1^{\mathsf{T}} v_1| \geq \sqrt{1 - \frac{\|P\|^2}{|\kappa_1 - \lambda_2|^2}}.$$

Proof. Take $i_0 = i_1 = 1$. By Theorem 37, $|\sin\theta(u_1, v_1)| \leq \frac{\|P\|}{|\kappa_1 - \lambda_2|}$. The corollary follows as $|u_1^{\mathsf{T}} v_1| = |\cos\theta(u_1, v_1)| = \sqrt{1 - \sin^2\theta(u_1, v_1)}$. □

Theorem 39 (Courant-Fischer-Weyl min-max principle)**.** *For every* $i = 1, 2, \ldots, n$*,*

$$\begin{aligned}
\lambda_i &= \min_{U \preceq \mathbb{R}^n} \left\{ \max_{x \in U} \left\{ \frac{x^{\mathsf{T}} B x}{x^{\mathsf{T}} x} : x \neq 0 \right\} : \dim(U) = i \right\} \\
&= \max_{U \preceq \mathbb{R}^n} \left\{ \min_{x \in U} \left\{ \frac{x^{\mathsf{T}} B x}{x^{\mathsf{T}} x} : x \neq 0 \right\} : \dim(U) = n - i + 1 \right\}.
\end{aligned}$$

Let C be an $m \times m$ Hermitian matrix with eigenvalues $\nu_1 \leq \nu_2 \leq \cdots \leq \nu_m$, where $m \leq n$.

Theorem 40 (Cauchy's interlacing theorem)**.** *If* $C = \Pi^* B \Pi$ *for an orthogonal projection* Π*, then for all* $i = 1, 2, \ldots, m$ *it holds that* $\lambda_i \leq \nu_i \leq \lambda_{n-m+i}$*.*

Theorem 41 (Rouché's theorem)**.** *Let* f *and* g *be two complex-valued functions that are holomorphic inside a region* R *with a closed simple contour* ∂R*. If for every* $x \in \partial R$ *we have that* $|g(x)| < |f(x)|$*, then* f *and* $f + g$ *have the same number of zeros inside* R*, counting multiplicities.*

Theorem 42 (Pan's Algorithm: Theorem 1.1 of [33])**.** *Given a monic degree* k *polynomial* $\hat{p}$ *with roots* $\rho_1, \ldots, \rho_k \in B(0,1)$ *and an accuracy parameter* $\gamma > 1$*, we can compute approximate roots* $\tilde{\rho}_1, \ldots, \tilde{\rho}_k$ *satisfying* $\|\rho - \tilde{\rho}\|_\infty \leq 2^{-\gamma}$ *in time* $O(k \log^2 k \cdot (\log\gamma + \log^2 k))$*.*

Corollary 43. *Let $q \in \mathbb{R}^{k+1}$ represent the polynomial $\hat{q}(z) = \prod_{i=1}^{k}(z - \theta_i)$ where $\theta_1, \ldots, \theta_k \in [0, 1]$ are ζ-separated, and let $r \in \mathbb{R}^{k+1}$ represent a polynomial of degree k with roots $\beta_1, \ldots, \beta_k$ satisfying $d(\theta, \beta) \leq \rho < \zeta/2$. For every $\varepsilon \in (0, \zeta/2 - \rho)$, we can reconstruct biases $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ satisfying $\left\|\tilde{\theta} - \theta\right\|_\infty \leq \rho + \varepsilon$ using $O(k \log^2 k \cdot (\log\log(1/\varepsilon) + \log^2 k))$ arithmetic operations.*

Proof. We will first approximate the roots $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ of $\hat{r}$ using Theorem 42. Since the roots of $\hat{r}$ are in $B\left(0, \frac{2k-1}{2k-2}\right)$ instead of $B(0, 1)$, we will actually find the roots of $\hat{r}\left(\frac{2k-2}{2k-1}z\right)$ and then multiply by $\frac{2k-1}{2k-2}$ to get the roots of $\hat{r}$ up to accuracy ε in time $O(k \log^2 k \cdot (\log\log(1/\varepsilon) + \log^2 k))$. (Notice that in order to get the desired accuracy we need to run Pan's algorithm to get the rescaled roots to within distance $\frac{2k-2}{2k-1} \cdot \varepsilon$; this doesn't matter for the purposes of runtime.)

Our output is $\tilde{\theta}_i \coloneqq \mathrm{Proj}_{[0,1]}(\tilde{\beta}_i)$ for $i = 1, \ldots, k$, where we label the roots $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ by the permutation achieving the matching distance, i.e., the ordering of coordinates so that $\|\theta - \beta\|_\infty = d(\theta, \beta)$. Now

$$\left|\theta_i - \tilde{\theta}_i\right| \leq \left|\theta_i - \Re(\tilde{\beta}_i)\right| \leq |\theta_i - \Re(\beta_i)| + \left|\Re(\beta_i) - \Re(\tilde{\beta}_i)\right| \leq |\theta_i - \beta_i| + \left|\beta_i - \tilde{\beta}_i\right| \leq \rho + \varepsilon.$$

□

Chapter 3

SUFFICIENT CONDITIONS FOR THE IDENTIFIABILITY OF MIXTURES OF PRODUCTS

[1] S. L. Gordon and L. J. Schulman, "Hadamard extensions and the identification of mixtures of product distributions," *IEEE Transactions on Information Theory*, vol. 68, no. 6, pp. 4085–4089, 2022. DOI: 10.1109/TIT.2022.3146630,

3.1 Introduction

The Hadamard product for row vectors $u = (u_1, \ldots, u_k)$, $v = (v_1, \ldots, v_k)$ is the mapping $\odot : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k$ given by

$$u \odot v := (u_1 v_1, \ldots, u_k v_k).$$

The identity for this product is the all-ones vector $\mathbb{1}$. We associate with vector v the linear operator $v_\odot = \mathrm{diag}(v)$, a $k \times k$ diagonal matrix, so that

$$u \cdot v_\odot = v \odot u.$$

Throughout this chapter $\mathbf{m}$ is a real matrix with row set $[n] := \{1, \ldots, n\}$ and column set $[k]$; write $\mathbf{m}_i$ for a row and $\mathbf{m}^j$ for a column.

As a matter of notation, for a matrix Q and nonempty sets R of rows and C of columns, let $Q|_R^C$ be the restriction of Q to those rows and columns (with either index omitted if all rows or columns are retained).

Definition 44. The Hadamard Extension of $\mathbf{m}$, written $\mathbb{H}(\mathbf{m})$, is the $2^n \times k$ matrix with rows $\mathbf{m}_S$ for all $S \subseteq [n]$, where, for $S = \{i_1, \ldots, i_\ell\}$, $\mathbf{m}_S = \mathbf{m}_{i_1} \odot \cdots \odot \mathbf{m}_{i_\ell}$; equivalently $\mathbf{m}_S^j = \prod_{i \in S} \mathbf{m}_i^j$. (In particular $\mathbf{m}_\emptyset = \mathbb{1}$.)

This construction originated recently in learning theory [20], [34] where it is arises naturally and unavoidably when we wish to perform source identification (i.e., parameter estimation) given data from a mixture (convex combination) of k product distributions on n binary random variables. We explain the connection further in Section 3.2. Motivated by this application, we are interested in the following two questions:

(1) If $\mathbb{H}(\mathbf{m})$ has full column rank, must there exist a subset R of the rows, of bounded size, such that $\mathbb{H}(\mathbf{m}|_R)$ has full column rank?

(2) In each row of $\mathbf{m}$, assign distinct colors to the distinct real values. Is there a condition on the coloring that ensures $\mathbb{H}(\mathbf{m})$ has full column rank?

In answer to the first question we show:

Theorem 45. *If $\mathbb{H}(\mathbf{m})$ has full column rank then there is a set R of no more than $k-1$ of the rows of $\mathbf{m}$, such that $\mathbb{H}(\mathbf{m}|_R)$ has full column rank.*

Considering the more combinatorial second question, observe that if $\mathbf{m}$ possesses two identical columns then the same is true of $\mathbb{H}(\mathbf{m})$, and so the latter cannot have full column rank. Extending this further, suppose there are three columns C in which only one row r has more than one color. Then $\operatorname{rowspace}\mathbb{H}(\mathbf{m}|^C)$ is spanned by $\mathbb{1}|^C$ and $r|^C$, so again $\mathbb{H}(\mathbf{m})$ cannot have full column rank. Motivated by these necessary conditions, set:

Definition 46. For a matrix Q let $\mathsf{NAE}(Q)$ be the set of nonconstant rows of Q (NAE="not all equal"); let $\varepsilon(Q|^C) = |\mathsf{NAE}(Q|^C)| - |C|$; and let $\overline{\varepsilon}(Q) = \min_{C\neq\emptyset}\varepsilon(Q|^C)$. If $\overline{\varepsilon}(Q) \geq -1$ we say Q satisfies the NAE condition.

In answer to the second question we have the following:

Theorem 47. *If $\mathbf{m}$ satisfies the NAE condition then*

(*a*) *There is a restriction of $\mathbf{m}$ to some $k-1$ rows R such that $\overline{\varepsilon}(\mathbf{m}|_R) = -1$.*

(*b*) *$\mathbb{H}(\mathbf{m})$ is full column rank.*

(*As a consequence also $\mathbb{H}(\mathbf{m}|_R)$ is full column rank.*)

Apparently the only well-known example of the NAE condition is when $\mathbf{m}$ contains $k-1$ rows which are identical and whose entries are all distinct. Then the vectors $\mathbf{m}_\emptyset, \mathbf{m}_{\{1\}}, \mathbf{m}_{\{1,2\}}, \ldots, \mathbf{m}_{\{1,\ldots,k-1\}}$ form a nonsingular Vandermonde matrix. This example shows that the bound of $k-1$ in (a) is best possible.

For another example in which the NAE condition ensures that $\operatorname{rank}\mathbb{H}(\mathbf{m}) = k$, take the $(k-1)$-row matrix with $\mathbf{m}_i^j = 1$ for $i \leq j$ and $\mathbf{m}_i^j = 1/2$ for $i > j$. Here the NAE condition is only minimally satisfied, in that for every $\ell \leq k$ there are ℓ columns C s.t. $\varepsilon(\mathbf{m}|^C) = -1$.

For $k > 3$ the NAE condition is no longer necessary in order that $\mathbb{H}(\mathbf{m})$ have full column rank. E.g., for $k = 2^\ell$, the $\ell \times k$ "Hamming matrix" $\mathbf{m}_i^j = (-1)^{j_i}$ where j is an ℓ-bit string $j = (j_1, \ldots, j_\ell)$, forms $\mathbb{H}(\mathbf{m}) =$ the Fourier transform for the group $(\mathbb{Z}/2)^\ell$ (often called a Walsh or Hadamard transform), which is invertible.

Furthermore, for $k \leq 2^\ell$, almost all (in the sense of Lebesgue measure) $\ell \times k$ matrices $\mathbf{m}$ form a full-column-rank $\mathbb{H}(\mathbf{m})$. (For $k = 2^\ell$ this is because $\det \mathbb{H}(\mathbf{m})$ is a polynomial in the entries of $\mathbf{m}$, and the Walsh example shows that this polynomial is nonzero. For $k < 2^\ell$, consideration of the same $2^\ell \times 2^\ell$ Walsh transform implies that there are some k rows of $\mathbb{H}(\mathbf{m})$ such that the determinant of the minor they form is a nonzero polynomial.) Despite this observation, the Vandermonde case, in which $k-1$ rows are required, is very typical, as it is what arises in $\mathbb{H}(\mathbf{m})$ for a mixture model of observables X_i that are iid conditional on a hidden variable. Another class of examples that is far from Lebesgue-typical, and furthermore also far from being "separated" (see next section), is this. There are two possible coins, with biases $p_1 \neq p_2$. A hidden variable H is sampled in $\{0, \ldots, k-1\}$, and then the process is that you observe the result of H independent tosses of coin 1, followed by $k - 1 - H$ independent tosses of coin 2. The NAE condition implies that here $\mathbb{H}(\mathbf{m})$ has full column rank. As a consequence (applying [34]) the following model is identifiable: a hidden H is sampled (from unknown prior) in $\{0, \ldots, k-1\}$, and then you observe the result of $2H$ independent tosses of coin 1 followed by $2k - 1 - 2H$ independent tosses of coin 2. A similar class of examples (sometimes identifiable but in general not) are the "subcube mixtures" studied in [20], where all coin biases must be one of $\{0, 1/2, 1\}$.

3.2 Motivation

Consider *observable* random variables $X_1, \ldots, X_n$ that are statistically independent conditional on H, a *hidden* or *latent* random variable H supported on $\{1, \ldots, k\}$. (See Figure 3.1 for an illustration.)

The most fundamental case is that the X_i are binary. Then we denote $\mathbf{m}_i^j = \Pr(X_i = 1 | H = j)$. The model parameters are $\mathbf{m}$ along with a probability distribution (the *mixture* distribution) $\pi = (\pi_1, \ldots, \pi_k)$ on H.

The study of finite mixture models was pioneered in the late 1800s in [35],

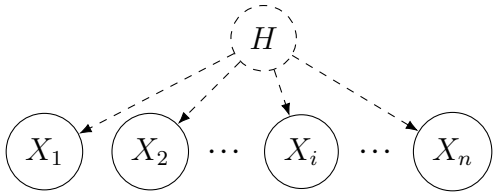

Figure 3.1: A Bayesian network diagram relating H and $X_1, \ldots, X_n$.

[36]. The problem of learning such distributions has drawn a great deal of attention. For surveys see, e.g., [37]–[40]. For some algorithmic papers on discrete-valued X_i, see [6], [9], [10], [15]–[20], [34]. The source identification (or parameter estimation) problem is that of computing $(\mathbf{m}, \pi)$ from the joint statistics of the X_i. Put another way, the problem is to invert the multilinear moment map

$$\begin{aligned} \mu : (\mathbf{m}, \pi) &\to \mathbb{R}^{2^{[n]}} \\ \mu(\mathbf{m}, \pi)_S &= \Pr(X_S = 1) \quad \text{where } S \subseteq [n],\ X_S = \prod\nolimits_{i \in S} X_i \\ &= \mathbf{m}_S \cdot \pi^\top. \end{aligned}$$

Since $\mathbf{m}_S^j = \Pr(X_S = 1 | H = j)$, this shows the essential role of $\mathbb{H}(\mathbf{m})$ in the mixture model.

Connection to $\operatorname{rank} \mathbb{H}(\mathbf{m})$

In general μ is not injective (even allowing for permutation among the values of π and columns of $\mathbf{m}$). For instance it is clearly not injective if $\mathbf{m}$ has two identical columns (unless π places no weight on those). More generally, and assuming all $\pi_j > 0$, it cannot be injective unless $\mathbb{H}(\mathbf{m})$ has full column rank. (Suppose $\alpha \in \mathbb{R}^k$ is nonzero s.t. $\mathbb{H}(\mathbf{m})\alpha = 0$. Since $(\mathbb{H}(\mathbf{m})\alpha)|_{\{\emptyset\}} = 0$, $\sum_j \alpha_j = 0$. So for sufficiently small $\delta > 0$, $\pi + \delta\alpha$ is a mixture distribution, distinct from π, with identical statistics.)

One sufficient condition for injectivity, due to [41], is that there be $2k - 1$ "separated" observables X_i. X_i is separated if all $\mathbf{m}_i^j$ are distinct, or in our terminology, if no color recurs in $\mathbf{m}_i$. (Further it is shown in [34], Theorem 1, that one can lower bound the distance between $\mu(\mathbf{m}, \pi)$ and any $\mu(\mathbf{m}', \pi')$ in terms of $\zeta = \min_i \min_{j \neq j'} |\mathbf{m}_i^j - \mathbf{m}_i^{j'}|$ and the distance between $(\mathbf{m}, \pi)$ and $(\mathbf{m}', \pi')$.) There are examples with $X_1, \ldots, X_{2k-1}$ where the mapping is injective but is no longer so if any single X_i is omitted [9].

A weaker and still sufficient condition for injectivity of μ, due to [34], is that for every $i \in [n]$ there exist two disjoint sets $A, B \subseteq [n] - \{i\}$ such that $\mathbb{H}(\mathbf{m}|_A)$ and $\mathbb{H}(\mathbf{m}|_B)$ have full column rank. (It is not known whether two disjoint such A, B are strictly necessary.)

Observable X_i with larger finite range

If each X_i can take on one of say L values, $\mathbf{m}$ can be considered as a nonnegative $n \times k \times L$ real array, with $\mathbf{m}^j_{i,\ell} = \Pr(X_i = \ell | H = j)$, $\sum_{\ell=1}^{L} \mathbf{m}^j_{i,\ell} = 1$; the multivariate moments are indexed not by sets S but by mappings $S : [n] \to [L]$, with $\mathbf{m}_S = \mathbf{m}_{S(1)} \odot \cdots \odot \mathbf{m}_{S(n)}$ and

$$\begin{aligned} \mu : (\mathbf{m}, \pi) &\to \mathbb{R}^{[L]^{[n]}} \\ \mu(\mathbf{m}, \pi)_S &= \Pr(X_S = 1) \qquad \text{where } X_S = \prod_{i=1}^{n} \delta_{X_i, S(i)} \\ &\qquad\qquad\qquad\qquad \text{(Kronecker delta)} \\ &= \mathbf{m}_S \cdot \pi^\top \end{aligned}$$

For any given k, if L is sufficiently large and $\mathbf{m}$ satisfies a certain nonsingularity condition, the mixture learning problem becomes easier; this insight is due to [6]. It will be interesting to explore what conditions exactly $\mathbf{m}$ must satisfy for identifiability (for positive π), for arbitrary L. But in this paper we study only the most extreme, and hardest for identification, case $L = 2$.

3.3 Some Theory for Hadamard Products, and a Proof of Theorem 45

For $v \in \mathbb{R}^k$ and U a subspace, extend the definition of $v_\odot$ to

$$v_\odot(U) = \{u \cdot v_\odot : u \in U\}$$

and introduce the notation

$$v_{\bar{\odot}}(U) = \operatorname{span}(U \cup v_\odot(U)).$$

We want to understand which subspaces U are invariant under $v_{\bar{\odot}}$. Let v have distinct values $\lambda_1 > \ldots > \lambda_\ell$ for $\ell \leq k$. Let the polynomials $p_{v,i}$ $(i = 1, \ldots, \ell)$ of degree $\ell - 1$ be the Lagrange interpolation polynomials for these values, so $p_{v,i}(\lambda_j) = \delta_{ij}$ (Kronecker delta). Let $B(v)$ denote the partition of $[k]$ into blocks $B(v)_{(i)} = \{j : v_j = \lambda_i\}$. Let $V_{(i)}$ be the space spanned by the elementary basis vectors in $B(v)_{(i)}$, and $P_{(i)}$ the projection onto $V_{(i)}$ w.r.t. the standard

inner product. Since $v_\odot$ is diagonal with entries λ_i in $B(v)_{(i)}$, we have the matrix equation

$$p_{v,i}(v_\odot) = P_{(i)}, \tag{3.1}$$

where $p_{v,i}$ is interpreted as a matrix polynomial. The collection of all linear combinations of the matrices $P_{(i)}$ is a commutative algebra, the $B(v)$ *projection algebra*, which we denote $A_{B(v)}$. The identity of the algebra is $I = \sum P_{(i)}$.

Definition 48. A subspace of $\mathbb{R}^k$ *respects* $B(v)$ if it has a basis in which each vector lies in some $V_{(i)}$.

For a subspace U we let $U^\perp$ be its orthogonal complement w.r.t. the standard inner product.

For U respecting $B(v)$ write $U = \mathrm{span}(\bigcup U_{(i)})$ for $U_{(i)} \subseteq V_{(i)}$. (Thus $U = \bigoplus U_{(i)}$ and $U_{(i)} = P_{(i)}U$.) Let $D_{(i)} = (U_{(i)})^\perp \cap V_{(i)}$. Then $(U_{(i)})^\perp = D_{(i)} \oplus \bigoplus_{j\neq i} V_{(j)}$.

Lemma 49. *A subspace $U^\perp$ respects $B(v)$ if U does.*

Proof. The subspaces of an inner product space form an orthocomplemented lattice in which the meet operation is intersection, and the negation operation is orthogonal complement. So for any subspaces W, W' we have De Morgan's law $(\mathrm{span}(W \cup W'))^\perp = W^\perp \cap W'^\perp$. Thus $U^\perp = \bigcap (U_{(i)})^\perp = \bigoplus D_{(i)}$. □

Lemma 50. *A subspace U respects $B(v)$ iff $U = \bigoplus (P_{(i)}U)$.*

Proof. ($\Leftarrow$): Because this gives an explicit representation of U as a direct sum of subspaces each restricted to some $V_{(i)}$.

($\Rightarrow$): By definition U is spanned by some collection of subspaces $V'_{(i)} \subseteq V_{(i)}$; since these subspaces are necessarily orthogonal, $U = \bigoplus V'_{(i)}$. Moreover, since $P_{(i)}$ annihilates $V_{(j)}, j \neq i$, and is the identity on $V_{(i)}$, it follows that each $V'_{(i)} = P_{(i)}U$. □

Theorem 51. *A subspace U is invariant under $v_{\bar\odot}$ iff U respects $B(v)$.*

Proof. ($\Leftarrow$): Let $w \in U$ and write $w = \sum w_i$ for $w_i \in U_{(i)}$. Then $v \odot w_i = \lambda_i w_i \in U_{(i)}$. So $v \odot w = \sum v \odot w_i \in \bigoplus U_{(i)} = U$.

($\Rightarrow$): If $U = v_{\bar\odot}(U)$ then these also equal $v_{\bar\odot}(v_{\bar\odot}(U))$, etc., so U is an invariant space of $A_{B(v)}$, meaning, $aU \subseteq U$ for any $a \in A_{B(v)}$. In particular, applying (3.1), this holds for $a = P_{(i)}$. So $U \supseteq \bigoplus (P_{(i)}U)$. On the other hand,

since $\sum P_{(i)} = I$, $U = (\sum P_{(i)})U \subseteq \bigoplus(P_{(i)}U)$. So $U = \bigoplus(P_{(i)}U)$. Now apply Lemma 50. □

The symbol $\subset$ is reserved for strict inclusion.

Lemma 52. *If* $R, T \subseteq [n]$ *and* $\operatorname{rowspace}\mathbb{H}(\mathbf{m}|_R) \subset \operatorname{rowspace}\mathbb{H}(\mathbf{m}|_{R\cup T})$, *then there is a row* $t \in T$ *such that* $\operatorname{rowspace}\mathbb{H}(\mathbf{m}|_R) \subset \operatorname{rowspace}\mathbb{H}(\mathbf{m}|_{R\cup\{t\}})$.

Proof. Without loss of generality R, T are disjoint. Let $T' \subseteq T$ be a smallest set s.t. $\exists R' \subseteq R$ s.t. $\mathbf{m}_{R'} \odot \mathbf{m}_{T'} \notin \operatorname{rowspace}\mathbb{H}(\mathbf{m}|_R)$. Select any $t \in T'$ and write $\mathbf{m}_{R'} \odot \mathbf{m}_{T'} = \mathbf{m}_{R'} \odot \mathbf{m}_{T'-\{t\}} \odot \mathbf{m}_t$. By minimality of T', $\mathbf{m}_{R'} \odot \mathbf{m}_{T'-\{t\}} \in \operatorname{rowspace}\mathbb{H}(\mathbf{m}|_R)$. But then $\mathbf{m}_{R'} \odot \mathbf{m}_{T'} \in \operatorname{rowspace}\mathbb{H}(\mathbf{m}|_{R\cup\{t\}})$, so $\operatorname{rowspace}\mathbb{H}(\mathbf{m}|_R) \subset \operatorname{rowspace}\mathbb{H}(\mathbf{m}|_{R\cup\{t\}})$. □

Proof of Theorem 45. This is now a consequence of Lemma 52. Start with $R = \emptyset$, and repeatedly use the Lemma to adjoin to R a row from $[n] \setminus R$ which will increase the rank of $\mathbb{H}(\mathbf{m}|_R)$ by at least 1. □

Remark

$\operatorname{rank}\mathbb{H}(\mathbf{m})$, along with a basis (using only rows of $\mathbb{H}(\mathbf{m})$) for $\operatorname{rowspace}\mathbb{H}(\mathbf{m})$, can be computed in time $O(nk^3)$ using Chen and Moitra's "GrowByOne" procedure [20]. For completeness here is a version of that procedure: For $\ell \geq 0$ let $W_\ell = \operatorname{span}(\mathbf{m}|_{[\ell]})$, and let $r_\ell = \operatorname{rank} W_\ell$. W_ℓ is spanned by some vectors $\mathbf{m}_{S_{\ell,1}}, \ldots, \mathbf{m}_{S_{\ell,r_\ell}}$, with all $S_{\ell,i} \subseteq [\ell]$, which we compute as follows. For $\ell = 0$ we have $r_0 = 1$, $S_{0,1} = \emptyset$. For $\ell > 1$ form the matrix with rows $\mathbf{m}_{S_{\ell-1,1}}, \ldots, \mathbf{m}_{S_{\ell-1,r_{\ell-1}}}$ followed by rows $\mathbf{m}_{S_{\ell-1,1}\cup\{\ell\}}, \ldots, \mathbf{m}_{S_{\ell-1,r_{\ell-1}}\cup\{\ell\}}$. Perform Gaussian elimination to zero-out all but $r_\ell - r_{\ell-1}$ of the second batch of rows. The first batch, together with the non-eliminated rows of the second batch, become $\mathbf{m}_{S_{\ell,1}}, \ldots, \mathbf{m}_{S_{\ell,r_\ell}}$.

3.4 Combinatorics of the NAE Condition: Proof of Theorem 47(a)

Recall we are to show: *If* $\overline{\varepsilon}(\mathbf{m}) \geq -1$ *then* $\mathbf{m}$ *has a restriction to some* $k-1$ *rows on which* $\overline{\varepsilon} = -1$.

Proof of Theorem 47(a). We induct on k. The (vacuous) base-case is $k = 1$.

For $k > 1$, we proceed by way of contradiction. Suppose the theorem fails for k, and let $\mathbf{m}$ be a k-column counterexample with the least possible number of rows, n. So $n > k - 1 \geq 1$. Necessarily every row of $\mathbf{m}$ is in $\mathsf{NAE}(\mathbf{m})$. Our

strategy is to show $\mathbf{m}$ has a restriction $\mathbf{m}'$ to $n-1$ rows, for which $\overline{\varepsilon}(\mathbf{m}') \geq -1$; this will imply a contradiction because, by minimality of the number of rows of $\mathbf{m}$, $\mathbf{m}'$ has a restriction to $k-1$ rows on which $\overline{\varepsilon} = -1$.

If $\overline{\varepsilon}(\mathbf{m}) \geq 0$ then we can remove any single row of $\mathbf{m}$ and still satisfy $\overline{\varepsilon} \geq -1$.

Otherwise, $\overline{\varepsilon}(\mathbf{m}) = -1$, so there is a nonempty S such that $|\,\mathsf{NAE}(\mathbf{m}|^S)| = |S| - 1$; choose a largest such S. It cannot be that $S = [k]$ (as then $n = k-1$). Arrange the rows $\mathsf{NAE}(\mathbf{m}|^S)$ as the bottom $|S|-1$ rows of the matrix. As discussed earlier, for the NAE condition one may regard the distinct real values in each row of $\mathbf{m}$ simply as distinct colors; relabel the colors in each row above $\mathsf{NAE}(\mathbf{m}|^S)$ so the color above S is called "white." (There need be no consistency among the real numbers called white in different rows.) See Fig. 3.2.

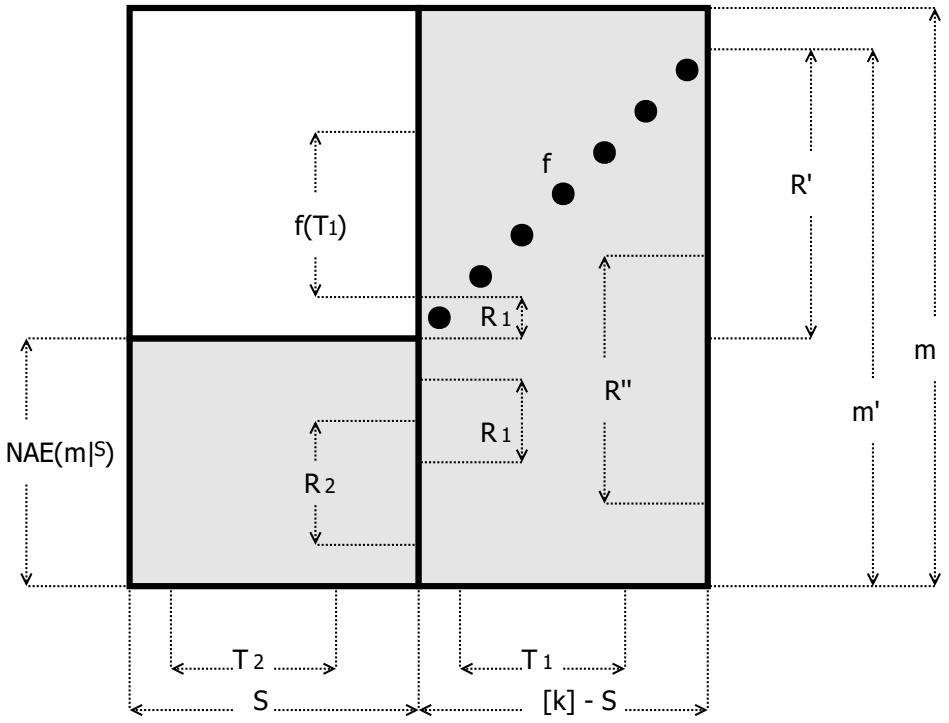

Figure 3.2: Argument for Theorem 47(a). Upper-left region is white. Entries $(t, f(t))$ (indicated with black dots) are not white.

Due to the maximality of $|S|$ and the fact that $\overline{\varepsilon}(\mathbf{m}) \geq -1$, there is no set of columns S' with $S \subset S'$ such that for some set of rows $A \subseteq [n] - \mathsf{NAE}(\mathbf{m}|^S)$, with $|A| = n - |S'| + 1$, $\mathbf{m}|_A^{S'}$ is all white. That is to say, if we form a bipartite graph on "right" vertices corresponding to the columns $[k] - S$, and "left"

vertices corresponding to the rows $[n] - \mathsf{NAE}(\mathbf{m}|^S)$, with non-white cells being edges, then any subset of the right vertices of size $\ell \geq 1$ has at least $\ell + 1$ neighbors within the left vertices.

By the induction on k (since $S \neq \emptyset$), for the set of columns $[k] - S$ there is a set R'' of $k - |S| - 1$ rows such that $\overline{\varepsilon}(\mathbf{m}|_{R''}^{[k]-S}) = -1$. Together with the rows of $\mathsf{NAE}(\mathbf{m}|^S)$ this amounts to at most $k - 2$ rows, so since $n \geq k$, we can find two rows outside this union; delete either one of them, leaving a matrix $\mathbf{m}'$ with $n - 1$ rows. This matrix has the rows $\mathsf{NAE}(\mathbf{m}|^S)$ at the bottom, and $n - |S|$ remaining rows which we call R'. The lemma will follow by showing that $\overline{\varepsilon}(\mathbf{m}') \geq -1$.

In $\mathbf{m}'$, the induced bipartite graph on right vertices $[k] - S$ and left vertices R' has the property that any right subset of size $\ell \geq 1$ has a neighborhood of size at least ℓ in R'. Applying Hall's Marriage Theorem, there is an injective $f : [k] - S \to R'$ employing only edges of the graph.

Now consider any nonempty set of columns T, and write it as $T = T_1 \cup T_2$ for $T_1 \subseteq [k] - S$ and $T_2 \subseteq S$. We need to show that $\varepsilon(\mathbf{m}'|^T) \geq -1$. Let $R_1 = \mathsf{NAE}(\mathbf{m}'|^{T_1}) \cap R''$ and $R_2 = \mathsf{NAE}(\mathbf{m}'|^{T_2})$. We have that $|R_1| \geq |T_1| - 1$ because $\overline{\varepsilon}(\mathbf{m}|_{R''}^{[k]-S}) = -1$. We further have that $|R_2| \geq |T_2| - 1$ because $\overline{\varepsilon}(\mathbf{m}) \geq -1$ and because $\mathsf{NAE}(\mathbf{m}|^{T_2}) \subseteq \mathsf{NAE}(\mathbf{m}|^S) = \mathsf{NAE}(\mathbf{m}'|^S)$, so no row of $\mathsf{NAE}(\mathbf{m}|^{T_2})$ has been removed in $\mathbf{m}'$.

If $T_2 = \emptyset$, the rows R_1 witness that $\varepsilon(\mathbf{m}'|^T) \geq -1$. Likewise if $T_1 = \emptyset$, the rows R_2 witness the same conclusion.

Lastly suppose both T_1 and T_2 are nonempty. Nonemptyness of T_2 gives $|\mathsf{NAE}(\mathbf{m}|^{T_2})| \geq |T_2| - 1$. Now use the matching f. The set of rows $f(T_1)$ lies in R' and is therefore disjoint from $\mathsf{NAE}(\mathbf{m}|^{T_2})$, which as noted is a subset of $\mathsf{NAE}(\mathbf{m}'|^S)$. Moreover since $T_2 \neq \emptyset$, every entry (t, j) for $t \in T_2, j \in R'$ is white. On the other hand due to the construction of f, for every $t \in T_1$ the entry $(t, f(t))$ is non-white. Therefore every row in $f(T_1)$ is in $\mathsf{NAE}(\mathbf{m}'|^{T_1 \cup T_2})$. So $|\mathsf{NAE}(\mathbf{m}'|^{T_1 \cup T_2})| \geq |T_2| - 1 + |T_1|$, which is to say $\varepsilon(\mathbf{m}'|^T) \geq -1$. Thus $\overline{\varepsilon}(\mathbf{m}') \geq -1$. □

3.5 From NAE to Rank: Proof of Theorem 47(b)

Recall we are to show: $\mathbb{H}(\mathbf{m})$ *has full column rank if* $\overline{\varepsilon}(\mathbf{m}) \geq -1$.

Proof of Theorem 47(b). The case $k = 1$ is trivial. Now suppose $k \geq 2$ and that Theorem 47(b) holds for all $k' < k$. Any constant rows of $\mathbf{m}$ affect neither the hypothesis nor the conclusion, so remove them, leaving $\mathbf{m}$ with at least $k-1$ rows. Now pick any set, C, of $k-1$ columns of $\mathbf{m}$. By Theorem 47(a) there are some $k-2$ rows of $\mathbf{m}$, call them R', on which $\overline{\varepsilon}(\mathbf{m}|^C_{R'}) = -1$. Let v be a row of $\mathbf{m}$ outside R'. Let R'' denote the set of rows of $\mathbf{m}$ other than v. Since R'' contains R', by induction $\dim \operatorname{rowspace} \mathbb{H}(\mathbf{m}|^C_{R''}) = k-1$. Therefore $U := \operatorname{rowspace} \mathbb{H}(\mathbf{m}|_{R''}) \subseteq \mathbb{R}^k$ is of dimension at least $k-1$. We claim now that $\dim v_{\bar{\odot}}(U) = k$. (Note that $v_{\bar{\odot}}(U) = \operatorname{rowspace} \mathbb{H}(\mathbf{m})$.)

Suppose to the contrary that $\dim v_{\bar{\odot}}(U) = k-1$. It must then be that $\dim U = k-1$ and $v_{\bar{\odot}}(U) = U$. So as proven in Theorem 51, U respects $B(v)$. Since v is nonconstant, $B(v)$ is a partition of $[k]$ into $\ell \geq 2$ nonempty blocks $B(v)_{(i)}$, and $U = \bigoplus_{i=1}^{\ell} U_{(i)}$ with $U_{(i)} = P_{(i)}U_{(i)}$. So there is some i_0 for which $U_{(i_0)} \subset V_{(i_0)}$; specifically, $U_{(i)} = V_{(i)}$ for all $i \neq i_0$, and $\dim U_{(i_0)} = \dim V_{(i_0)} - 1$. Since $|B(v)_{(i_0)}| < k$, we know by induction that $P_{(i_0)} \operatorname{rowspace} \mathbb{H}(\mathbf{m}) = V_{(i_0)}$. But since $\operatorname{rowspace} \mathbb{H}(\mathbf{m}) = v_{\bar{\odot}}(U) = U$, this means that $P_{(i_0)}U = V_{(i_0)}$. Contradiction. □

Chapter 4

SOURCE IDENTIFICATION FOR MIXTURES OF PRODUCTS

[1] S. Gordon, B. H. Mazaheri, Y. Rabani, and L. Schulman, "Source identification for mixtures of product distributions," in *Proceedings of Thirty Fourth Conference on Learning Theory*, M. Belkin and S. Kpotufe, Eds., ser. Proceedings of Machine Learning Research, vol. 134, PMLR, Aug. 2021, pp. 2193–2216. [Online]. Available: https://proceedings.mlr.press/v134/gordon21a.html,

4.1 Introduction

A k-MixProd is a mixture (that is, a convex combination) of k product distributions on a set of n random variables $\mathcal{X}$. In the notation of Bayesian networks, this situation is represented graphically by a single unobservable random variable U with edges to each of the variables $X_i \in \mathcal{X}$. U is referred to as a "confounding" variable with range $1, \ldots, k$ and the variables in $\mathcal{X}$ are referred to as the "observables." The main complexity parameter of the problem is k, the number of mixture constituents or "sources." See Fig. 4.1.

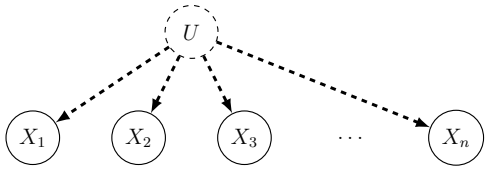

Figure 4.1: Graphical depiction of a k-MixProd

In this chapter we study the *identification* problem for k-MixProds. Specifically, given a joint distribution $\mathcal{P}$ on the variables (vertices), recover up to small statistical error (a) the mixture weights, up to a permutation of the constituents, and (b) the conditional distribution of each vertex X within each mixture source.

Problem Statement and Notation.

The hidden variable U ranges in $[k] = \{1, \ldots, k\}$, with unknown distribution $\pi_j := \Pr(U = j)$. There are n binary observable variables $X_1, X_2, \ldots, X_n$; their distributions are given by the rows $\mathbf{m}_i$ of an unknown matrix $n \times k$ matrix

$\mathbf{m}$, with the meaning that $\mathbf{m}_{ij} := \Pr[X_i = 1 \mid U = j]$. Thus, the identification problem is to recover $(\pi_j)_{j\in[k]}$ and $(\mathbf{m}_{ij})_{i\in[n],j\in[k]}$ (up to permuting the set $[k]$) to within high accuracy.

Identification is not always possible, as there can be disparate models sharing the same distribution on the observables. To give sufficient conditions for identifiability, we need the following definition, which allows us to quantify the degree to which an observable variable has distinct behavior in each of the different mixture constituents.

Definition 53. An observable X_i is ζ-*separated* if $\min_{j\neq j'} |\mathbf{m}_{ij} - \mathbf{m}_{ij'}| > \zeta$.

It is known that a k-MixProd is identifiable (from perfect statistics) if it has at least $2k-1$ 0-separated observables [42] and $\pi_{\min} > 0$. In general (although not for every $\mathbf{m}$), $2k-1$ observables are also necessary [9]. Moreover some separation assumption is certainly necessary for identification: e.g., a mixture of several identical sources generates the same statistics as a single source.

Clearly, a *sample size* bound requires a further, quantitative assumption on separation; this is the role of the parameter ζ in our work. For our near-optimal running time and sample complexity guarantee, we also require a slightly larger-than-minimal number of observables: $3k-3$ rather than $2k-1$. (A similar algorithm works with just $2k-1$ ζ-separated observables, and its post-sampling runtime is about the same, but its sample size requirement is larger.)

It should be noted that the post-sampling runtime, although exponential in k, is hardly affected by ζ. The bottleneck resource is sample size.

Main Result

Theorem 54. *The algorithm given later w.h.p. identifies a k-MixProd on $n \geq 3k-3$ binary observables, each of which is ζ-separated, with runtime* (*and, a fortiori, sample complexity*)

$$(1/\zeta)^{O(k\log k)}(n\log n)(1/\min \pi_i)^{O(\log k)}(1/\varepsilon)^2.$$

and w.h.p. (*over the distribution of samples*) *computes $\tilde{\pi}$ and $\tilde{\mathbf{m}}$ such that $\max_i |\tilde{\pi}_i - \pi_i| \leq \varepsilon$ and $\max_{ij} |\tilde{\mathbf{m}}_i^j - \mathbf{m}_i^j| \leq \varepsilon$.*

It is actually sufficient that there merely *exist* a subset of $3k-3$ ζ-separated random variables. Under this weaker assumption, the runtime above is multiplied by $n^{O(k)}$ (the sample size is unaffected).

Discussion

The k-MixProd problem for finite-range observables has been studied for nearly 30 years [15]–[19], [42]–[44] There are two versions of the problem: (1) *Learning* the model, namely, producing any model consistent with (or close to) the observed statistics; (2) *Identifying* the model, namely, producing the true model (or one close to it). Any algorithm for "learning" will also of course achieve identification if it happens that the model is uniquely specified by the statistics, but the algorithm may not, and existing algorithms do not, provide a certificate of uniqueness under such conditions.

Our result is concerned entirely with identification. Since our algorithm computes, as part of its operation, the condition number of the matrices that it uses for inverting the model→statistics mapping, it will only return an output under conditions which ensure the output is indeed unique (to within the allowed ε error). Hence, as compared with a "learning" algorithm, we are effectively providing a stronger output guarantee under stronger conditions on the input.

One of the chief motivations for our work is the characterization of *interventional distributions* in Bayesian networks (causal DAGs). Causal inference is a sizable and growing field; in that setting, we want to understand the distribution on some "downstream" variable(s) that will result if we disconnect a specified "intervention variable" from its parents in the DAG, and instead assign it to a chosen value. This task is easy when all variables in the network are observable, but challenging when there are latent or "hidden" variables whose statistics we cannot observe. In the most extreme case there may be a latent variable U which can directly affect all observables. It turns out, as we show in separate work, that it is possible to solve that problem when U takes on a bounded number k of values. In that work, a key subroutine has to solve k-MixProd problems. The role of the present work is to provide and analyze the needed algorithm for that problem.

We should emphasize that "learning" the model, without "identification," is (so far as we know) useless for causal inference applications. Fortunately,

we show that when identification is possible, it can also be performed much faster than the best current results for "learning" (which suffer a k^3 in the exponent, as compared with our $k \log k$).

Of course it remains of great interest to ask whether a similar runtime can be achieved for learning. Our algorithm, in any event, does not achieve this.

Comments on the assumptions in Theorem 54.

1. *The observable variables are binary.* The binary case is actually the most difficult. (The problem also becomes significantly easier when each variable X_i takes on at least k values and the k distributions on X_i span a k-dimensional space. This kind of assumption was first to our knowledge used in, in a somewhat different setting, in [45].)
2. *The mixture is ζ-separated.* A separation condition of some kind, on at least some of the variables, is necessary, since we certainly cannot identify the distribution of the latent variable if it does not have sufficient effect upon the observable variables. The 0-separation condition fails only on a set of Lebesgue measure 0 in the parameter space. Likewise for small ζ, most distributions are ζ-separated. The value of ζ-separation will be, as we show below, that it ensure that matrices representing the parameters for each source are well-conditioned. ζ-separation is not always a necessary condition; characterizing necessary conditions is a difficult open question, tackled in part in [46], where it is shown, essentially, that some "batching" of not-fully-separated variables is also sufficient to imply invertibility of the relevant matrices (however, no quantitative bounds on condition number are known).
3. *The required value of n.* Depending on the actual distributions (as expressed by $\mathbf{m}$), the number of ζ-separated variables n required to identify the source may be anywhere from $\lg n$ to $2k-1$. Less than $\lg n$ is not possible: in that case a k-MixProd can be *any* probability distribution on $\mathcal{X}$, and it is not hard to see that parameters are not unique.

Prior work. In [19] a seminal learning algorithm for k-MixProd was given. Its running time, for mixtures on n binary variables (n suff. large), is $n^{O(k^3)}$. This was improved in [43] to $k^{O(k^3)}n^{O(k^2)}$. The most recent algorithm [44] identifies a mixture of k product distributions on at least $3k-3$ ζ-separated variables in time $2^{O(k^2)}n$. In this work we improve the latter results, giving

upper bounds with exponent $O(k \log k)$ rather than $O(k^2)$. This is nearly optimal, as it is known that even in the special case that the observables are known to be iid conditional on U—call this the k-MixIID problem—exponent $\Omega(k)$ is unavoidable [9]. Moreover $n \geq 2k - 1$ observable variables are generally necessary. (There are interesting special cases in which better efficiency is possible. [43] study "mixtures of sub-cubes," i.e., k-MixProd with all bit probabilities in $\{0, \frac{1}{2}, 1\}$. For this problem they achieve complexity $n^{O(\log k)}$.)

It turns out that the k-MixIID problem plays a special role at the base of a tower of reductions. On the one hand it can be solved using a relatively simple, two-century-old method of Prony [7], which connects it to the classical Hausdorff moment problem; it can also be solved by the Matrix Pencil Method [3], [9], [10], [47]. Both analyses are quite recent. In turn, the k-MixProd problem is solved by a nontrivial reduction to k-MixIID, first with $\exp(O(k^2))$ complexity [44] and here with $\exp(O(k \log k))$ complexity. In forthcoming work the authors have shown in turn how to reduce to k-MixProd from the yet more general k-MixBND problem, in which the input distribution is a k-component mixture over distributions consistent with a (known) Bayesian Network. The dominant term in the complexity of that algorithm is that of the k-MixProd instances to which the problem reduces. Hence the key role played by the improved sample- and run-time complexities of the present work.

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 the discrete case. It is necessary however to point out a fundamental difference between the types of problems. In general, in a k-MixIID problem, a mixture source is chosen from distribution π), and then we sample some n independent samples *from that source*. In almost every parametric scenario (think e.g., of a mixture of k unit-variance gaussians 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. To see this, consider an instance of 2-MixIID for binary variables with two equiprobable sources (i.e., $\pi_1 = \pi_2 = 1/2$), one of which has $\Pr(X = 1 \mid U = 0) = \frac{3}{4}$ and the other $\Pr(X = 1 \mid U = 1) = \frac{1}{4}$. With only a single sample from each source, it is impossible to distinguish between a

mixture in which $\Pr(X = 1 \mid U = 0) = 1$ and the other $\Pr(X = 1 \mid U = 1) = 0$. More specifically, it is only possible to identify the *mean* of these two parameters. With access to multiple samples from the *same* source, however, we can begin to infer the parameters associated with each mixture component. For the k-MixIID problem for binary rv's, it is known that $n = 2k - 1$ such samples [9] are needed. (This value is called there the threshold "aperture" of the problem.)

Unlike k-MixIID, the k-MixProd problem does not allow us access to multiple samples of iid variables from a source. However it does give access to multiple variables which are independent conditional on the source. An approach introduced in [44] is to create synthetic copies of a single variable via linear combinations of the other variables; since we need to modify the method, we describe below how this is done. This approach enables a reduction from the k-MixProd problem to the k-MixIID problem.

Organization. Section 4.2 sets up the key mathematical objects needed for our work. Section 4.3 describes the algorithm. The algorithm is similar to that in [44], differing in one crucial way which will be described. Along with the algorithm pointers are provided to the main steps of the analysis in Appendix 4.5. The part of this work which is a full departure from the previous literature is the condition number bound in Section 4.4, in which the condition number of a key linear mapping (namely, the matrix of multilinear moments of the distribution on observables), is bounded through a novel argument characterizing the images of rank-1 tensors under mapping by Hadamard extensions.

4.2 Preliminaries

Hadamard Extensions

As in earlier works [43], [44], the algorithm and its analysis make extensive use of the Hadamard extension of a matrix $\mathbf{m}$.

Definition 55. Given a matrix $\mathbf{m}$ of any dimensions, let $\mathbf{m}_{i*}$ denote the ith row of $\mathbf{m}$ and $\mathbf{m}_{*j}$ the jth column of $\mathbf{m}$. Where clear from context we write $\mathbf{m}_i$ instead of $\mathbf{m}_{i*}$.

Definition 56 (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 linear operator $v_\odot = \mathrm{diag}(v)$, the Hadamard product is $u \odot v = u \cdot v_\odot$. The identity element for the Hadamard product is the all-ones vector $\mathbb{1}$. The Hadamard product of u with itself k times is written $u^{\odot k}$.

Definition 57 (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 $\mathbb{H}(\mathbf{n})^j_S = \prod_{i\in S} \mathbf{n}^j_i$. In particular $\mathbb{H}(\mathbf{n})_\emptyset = \mathbb{1}$, and for all $i \in [n]$, $\mathbb{H}(\mathbf{n})_{\{i\}} = \mathbf{n}_i$.

Notation for subsets and collections of subsets We will reserve calligraphic fonts for collections of subsets of $[n]$, i.e., $\mathcal{S} \subseteq 2^{[n]}$. In contrast, the same variable in a standard math font will represent a subset of $[n]$, i.e., $S \subseteq [n]$.

Definition 58 (Sum of two collections)**.** The *sum of two collections* of subsets $\mathcal{X}, \mathcal{Y} \subseteq 2^{[n]}$ is defined as $\mathcal{X} + \mathcal{Y} \coloneqq \{X \cup Y : X \in \mathcal{X}, Y \in \mathcal{Y}\}$.

Definition 59 (Kronecker product of vectors)**.** Let $S, T \subset [n]$ be two disjoint sets. Let x and y, respectively, be vectors indexed by the subsets in $\mathcal{X} = 2^S$ and $\mathcal{Y} = 2^T$, respectively. Then the *Kronecker product* $x \otimes y$ is the vector indexed by the subsets in $\mathcal{X} + \mathcal{Y}$ given by

$$(x \otimes y)_{X\cup Y} \coloneqq x_X y_Y, \text{ for all } X \in \mathcal{X} \text{ and } Y \in \mathcal{Y}.$$

Note that $\mathcal{X} + \mathcal{Y} = 2^{S\cup T}$. (Notice that every set $Z \in \mathcal{X} + \mathcal{Y}$ can be written uniquely as $Z = X \cup Y$ for $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. We don't define the Kronecker product when $\mathcal{X} \cap \mathcal{Y} \neq \emptyset$.)

Definition 60 (Kronecker product of vectors — alternate definition)**.** Let $S, T \subset [n]$ be two disjoint sets. Let $x \in \mathbb{R}^{2^S}$ and $y \in \mathbb{R}^{2^T}$, which is to say, x has a coordinate x_R for every $R \subseteq S$, and similarly for y. Then the *Kronecker product* $x \otimes y$ is the vector with coordinates x_R for every $R \subseteq S \cup T$, given by

$$(x \otimes y)_R \coloneqq x_{R\cap S} y_{R\cap T}.$$

In the algorithm and analysis we will make heavy use of the singular values of a matrix $\mathbf{M}$, which we will denote $\sigma_{\max}(\mathbf{M}) = \sigma_1(\mathbf{M}) \geq \sigma_2(\mathbf{M}) \geq \cdots$.

When $\left\{v^{(i)}\right\}_{i=1}^k$ is a set of row vectors, we will write $(v^{(1)}; v^{(2)}; \ldots; v^{(k)})$ for the matrix obtained by vertically concatenating each row vector in order.

Finally, for a matrix $\mathbf{M}$, $\mathbf{M}^+$ will denote the Moore-Penrose inverse (i.e., the pseudoinverse), given by $(\mathbf{M}^\mathsf{T}\mathbf{M})^{-1}\mathbf{M}^\mathsf{T}$ in the case that $\mathbf{M}$ has full column rank.

Definition 61 (Column-wise Khatri-Rao product of matrices)**.** The *column-wise Khatri-Rao product* of matrices $\mathbf{A} \in \mathbb{R}^{n\times k}, \mathbf{B} \in \mathbb{R}^{m\times k}$ is denoted $\mathbf{A} * \mathbf{B}$; $\mathbf{A} * \mathbf{B}$ has dimensions $nm \times k$ with row ij given by $\mathbf{A}_{i*} \odot \mathbf{B}_{j*}$.

Definition 62 (Vandermonde matrix)**.** The Vandermonde matrix $\mathrm{Vdm}(\mathbf{m}) \in \mathbb{R}^{k\times k}$ associated with a vector $\mathbf{m} \in \mathbb{R}^k$ has entries $\mathrm{Vdm}(\mathbf{m})_{ji} = \mathbf{m}_i^j$ for $j \in \{0, 1, \ldots, k-1\}$ and $i \in \{1, 2, \ldots, k\}$.

Multilinear Moments

Using the notation from Sec. 4.1, observe that the expectation of each X_i is given by $\mathbf{m}_i\pi^\mathsf{T}$. The model parameters are $(\pi_j)_{j\in[k]}$ and $(\mathbf{m}_{ij})_{i\in[n],j\in[k]}$ (up to a permutation of the range of U).

When $k > 1$, observable variables X_i and $X_{i'}$ will not in general be independent. We will make use of information about the dependencies between observables through measurements of $\mathbb{E}[X_iX_{i'}]$. Generalizing this, for any $S \subseteq [n]$, we will define the random variable $X_S \coloneqq \prod_{i\in S} X_i$. The data we obtain from our samples will be estimates of $\mathbb{E}[X_S]$ for different subsets $S \subseteq [n]$. These are known as the *multilinear moments* of the distribution, since they are multilinear polynomials in the parameters $\mathbf{m}_{ij}$.

Our goal will be to eventually compute power moments for the single variable X_1, i.e. the values $\mathbb{E}[X_1X_1^{(2)}X_1^{(3)}\cdots X_1^{(\ell)}]$ for $\ell \le 2k$ where each $X_1^{(j)}$ is a copy of X_1 that is iid conditioned on U. Such copies are not guaranteed to exist among the X_i, but we can still define (and our algorithm will compute) the power moments. The ℓth power moment is defined as $\mathbb{E}[X_1^{\odot\ell}]$, the expectation of the product of ℓ copies of X_1 that are iid conditioned on U.

Now $\mathbb{E}[X_S] = \sum_j \pi_j \prod_{i\in S} \mathbf{m}_{ij}$, or equivalently, $\mathbb{E}[X_S] = (m_{i_1} \odot m_{i_2} \odot \cdots \odot m_{i_s})\pi^\mathsf{T}$ where $S = \{i_1, i_2, \ldots, i_s\}$.

If we replace $\mathbf{m}$ with the Hadamard extension $\mathbf{M} \coloneqq \mathbb{H}(\mathbf{m})$, we can simplify the above equation to $\mathbb{E}[X_S] = \mathbf{M}_S\pi^\mathsf{T}$.

We can also write power moments in terms of Hadamard products. In particular, we have $\mathbb{E}[X_i^{\odot\ell}] = \sum_j \pi_j \mathbf{m}_{ij}^\ell = \mathbf{m}_i^{\odot\ell}\pi^\mathsf{T}$.

Observe that source identification is not possible if $\mathbf{M}$ has less than full column rank, i.e., $\operatorname{rank}\mathbf{M} < k$, as then the mixing weights cannot be unique.

For a collection of subsets $\mathcal{S} \subseteq 2^{[n]}$, let $\mathbf{M}[\mathcal{S}]$ denote the restriction of $\mathbf{M}$ to the rows $\mathbf{M}_S, S \in \mathcal{S}$. E.g., $\mathbf{M} = \mathbf{M}[2^{[n]}]$.

If we have non-overlapping collections of subsets $\mathcal{A}$ and $\mathcal{B}$, we can build a matrix of multilinear moments as follows:

Definition 63 (Observable matrices)**.** For collections $\mathcal{A}, \mathcal{B} \subseteq 2^{[n]}$ with $\mathcal{A} \subseteq 2^S$ and $\mathcal{B} \subseteq 2^T$ for disjoint $S, T \subseteq [n]$, we define the matrix

$$\mathbf{C}_{\mathcal{B},\mathcal{A}} \coloneqq \mathbf{M}[\mathcal{B}]\pi_{\odot}\mathbf{M}[\mathcal{A}]^{\mathsf{T}}.$$

For any $A \in \mathcal{A}, B \in \mathcal{B}$, we have

$$(\mathbf{C}_{\mathcal{B},\mathcal{A}})_{B,A} = \mathbf{M}_B\pi_{\odot}\mathbf{M}_A^{\mathsf{T}}$$

which is the multilinear moment $\mathbb{E}[X_{A\cup B}]$ and therefore observable.

The Empirical Multi-linear Moments For a finite sample drawn from the model, we let $\tilde{\mathrm{g}}(S)$ be the empirical estimate of $\mathbb{E}[X_S]$, i.e., the fraction of samples for which $\prod_{i\in S} X_i = 1$. These $\tilde{\mathrm{g}}(S)$ for $S \subseteq [n]$ are the complete list of "observables" of the model. Each converges, in the infinite-sample limit, to the value $\mathrm{g}(S) \coloneqq \mathbb{E}[X_S]$,

$$\mathrm{g}(S) = \mathbf{M}_S\pi^{\mathsf{T}} = \mathbf{M}_S\pi_{\odot}\mathbb{1}^{\mathsf{T}}.$$

Our algorithm will heavily utilize (the empirical estimates of) matrices of observable moments.

4.3 The Algorithm

The algorithm appears below as Algorithm 3. It uses disjoint sets $S, T, T' \subseteq [n]$ of ζ-separated variables, with $|S| = |T| = |T'| = k-1$ and $1 \in T$, and takes as input the desired accuracy ε. We will use tildes in the algorithm to denote the empirical version of quantities we define with respect to exact moments. To avoid clutter however we will describe most steps of the algorithm as though we had access to the exact moments we need (i.e. $\mathrm{g}(S)$ instead of $\tilde{\mathrm{g}}(S)$). In the actual algorithm we use our empirical estimates of the moments.

Definition 64. Define the following collections from S, T, T':

$$\mathcal{A} \coloneqq 2^S, \quad \mathcal{B} \coloneqq 2^T, \quad \mathcal{B}' \coloneqq 2^{T'}.$$

We introduce shorthand for the corresponding submatrices of $\mathbf{M}$:

$$\mathbf{A} \coloneqq \mathbf{M}[\mathcal{A}], \quad \mathbf{B} \coloneqq \mathbf{M}[\mathcal{B}], \quad \mathbf{B}' \coloneqq \mathbf{M}[\mathcal{B}].$$

An essential difference between this algorithm and the one given in [44] is that our $\mathcal{A}, \mathcal{B}, \mathcal{B}'$ comprise of the *entirety* of 2^S, 2^T, and $2^{T'}$. Consequently, we use the entire $2^k \times 2^k$ matrix for $\mathbf{C}_{\mathcal{A}\mathcal{B}}$ rather than performing an expensive $2^{O(k^2)}$ search within it for a well-conditioned $k \times k$ submatrix. In the earlier paper this search cost was dominated by other costs, but that is no longer the case here, due to the stronger bounds we provide (coming from Section 4.4).

Our goal will be to compute the power moments $\mathbb{E}[X_1^{\odot \ell}]$. To do this, we will alternate between computing vectors of moments v_i, and vectors of coefficients u_i, u_i', defined below:

Definition 65. The sequence of vectors $v_1, \ldots, v_{2k} \in \mathbb{R}^{\mathcal{A}}$ is given by

$$(v_i)_A \coloneqq \mathbb{E}[X_A X_1^{\odot i}]$$

for each $A \in \mathcal{A}$.

In particular, $(v_i)_\emptyset = \mathbb{E}[X_1^{\odot i}]$.

Proposition 66. $v_i = m_1^{\odot i} \pi_\odot \mathbf{A}$.

That is, each entry of the vector v_i will be a *mixed* moment, with a multilinear part given by X_A and a power moment part given by $X_1^{\odot \ell}$. Our algorithms begins by computing $\tilde{v}_0$ and $\tilde{v}_1$, the empirical counterparts to v_0 and v_1, respectively. These empirical counterparts are available directly as empirical multilinear moments.

Definition 67. The sequence of vectors $u_1, \ldots, u_{2k} \in \mathbb{R}^{\mathcal{B}}$ is given by the equations

$$u_i \mathbf{M}[\mathcal{B}] \coloneqq m_1^{\odot i}.$$

The sequence of vectors $u_1', \ldots, u_{2k}' \in \mathbb{R}^{\mathcal{B}'}$ is defined analogously with $\mathcal{B}'$ replacing $\mathcal{B}$ everywhere.

Proposition 68. *$u_i = v_i \mathbf{C}^+_{\mathcal{B}\mathcal{A}}$ and $u'_i = v_i \mathbf{C}^+_{\mathcal{B}'\mathcal{A}}$.*

That is, u_i is the vector of coefficients of linear combinations of parameter vectors from $\mathbf{M}[\mathcal{B}]$ (or $\mathbf{M}[\mathcal{B}']$) needed to create the parameter vector for the variable $X_1^{\odot i}$.

Once we have computed $v_1, v_2, \ldots, v_{2k}$, we can take the entries $(v_i)_\emptyset = \mathbb{E}[X_1^{\odot i}]$ to obtain the first $2k$ power moments for X_1, which we can then pass as input to an algorithm for learning k-MixIID.

At this point, we will have recovered $\mathbf{m}_1$ and π and we can use these to recover the remaining parameters with some more straightforward linear algebra.

Empirical counterparts In the pseudocode for Algorithm 3, we will exclusively be working with empirically computed (hence, approximate) versions of the quantities above. The empirical counterpart to any quantity is distinguished by a tilde, e.g., $\tilde{\mathbf{A}}$ is the empirical counterpart to $\mathbf{A}$.

The Steps of the Algorithm

We now outline what each non-trivial line of the algorithm accomplishes.

- Line 1: As stated in Theorem 54, we write the algorithm under the assumption that all observable bits are ζ-separated (and consequently in this line, S, T, T' are arbitrary disjoint sets of size $k-1$). This assumption is not strictly necessary; all we actually need is that the disjoint sets S, T, T' create $\hat{\mathbf{C}}_{\mathcal{B}\mathcal{A}}$ and $\hat{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}$ with σ_k above the threshold in Line 6, something that is guaranteed if S, T, T' are of size $k-1$ and consist of ζ-separated variables, but may also hold without either of these conditions. In any case, provided one assumes that there exist some $3k-3$ ζ-separated variables, an exhaustive search running in time $n^{O(k)}$ will suffice to find S, T, T' passing the threshold in Line 6.

- Line 3: For the algorithm outlined here to produce $\mathbf{m}$ and π within additive error ε (in each coordinate), we shall need the $\tilde{\mathrm{g}}(S)$ that we use to be accurate to within $\varepsilon(\min \pi_i)^{O(\log k)}\zeta^{O(k \log k)}$. This can be achieved with high probability using a sample of size

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

input: disjoint sets of $\zeta-$separated S, T, T' of size $k-1$ each, with $1 \in T$, as well as the desired accuracy ε.

$\mathcal{A} \leftarrow 2^S, \mathcal{B} \leftarrow 2^T, \mathcal{B}' \leftarrow 2^{T'}$.

Use $(1/\varepsilon)^2(1/\min \pi_i)^{O(\log k)}(1/\zeta)^{O(k \log k)}$ samples to estimate $\{\tilde{\mathrm{g}}(A \cup B \cup B')\}_{A \in \mathcal{A}, B \in \mathcal{B}, B' \in \mathcal{B}'}$.

Construct $\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}$, $\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}$, and $\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$ from the empirical moments.

Compute $\hat{\mathbf{C}}_{\mathcal{B}\mathcal{A}}$ and $\hat{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}$ by taking the SVD of $\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}$ and $\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}$ and truncating to the first k singular values.

if $\min\{\sigma_k(\hat{\mathbf{C}}_{\mathcal{B}\mathcal{A}}), \sigma_k(\hat{\mathbf{C}}_{\mathcal{B}'\mathcal{A}})\} < \pi_{\min}\zeta^{O(k)}$ **then** terminate

$\tilde{v}_0 \leftarrow (\tilde{\mathrm{g}}(A))_{A \in \mathcal{A}}$

$\tilde{v}_1 \leftarrow (\tilde{\mathrm{g}}(A \cup \{1\}))_{A \in \mathcal{A}}$

$\tilde{u}_1 \leftarrow \tilde{v}_1 \hat{\mathbf{C}}^{+}_{\mathcal{B}\mathcal{A}}$

$\tilde{u}'_1 \leftarrow \tilde{v}_1 \hat{\mathbf{C}}^{+}_{\mathcal{B}'\mathcal{A}}$

for $i = 1, \ldots, 1 + \lg k$

  for $j = 1, \ldots, 2^{i-1}$

    $\tilde{v}_{2^{i-1}+j} \leftarrow (\tilde{u}_j \otimes \tilde{u}'_{2^{i-1}})\hat{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$

    $\tilde{u}_{2^{i-1}+j} \leftarrow \tilde{v}_{2^{i-1}+j}\hat{\mathbf{C}}^{+}_{\mathcal{B}\mathcal{A}}$

  $\tilde{u}'_{2^i} \leftarrow \tilde{v}_{2^i}\hat{\mathbf{C}}^{+}_{\mathcal{B}'\mathcal{A}}$.

Let $\mathcal{H}_{k+1}$ be the $(k+1) \times (k+1)$ Hankel matrix with entries given by $[\mathcal{H}_{k+1}]_{i,j} = (\tilde{v}_{i+j})_{\{1\}}$

if the second$-$smallest eigenvalue of $\mathcal{H}_{k+1}$ is below $\pi_{\min}\zeta^{O(k)}$ **then** terminate

$\tilde{\mathbf{m}}_1, \tilde{\pi} \leftarrow$ LearnPowerDistribution$(\mathcal{H}_{k+1})$.

$\tilde{V} \leftarrow (\tilde{v}_0; \ldots; \tilde{v}_{k-1})$

$\mathrm{Vdm}(\tilde{\mathbf{m}}_1) \leftarrow (\tilde{\mathbf{m}}_1^{\odot 0}; \ldots; \tilde{\mathbf{m}}_1^{\odot(k-1)})$

$\tilde{\mathbf{A}}^{\mathsf{T}} \leftarrow \tilde{\pi}_{\odot}^{-1}(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\tilde{V}$

$\tilde{\mathbf{B}} \leftarrow \hat{\mathbf{C}}_{\mathcal{B}\mathcal{A}}(\tilde{\mathbf{A}}^{\mathsf{T}})^{+}\pi_{\odot}^{-1}$.

for every $i \in [n] \setminus S$, $\tilde{\mathbf{m}}_i \leftarrow ((\tilde{\mathrm{g}}(R \cup \{i\}))_{R \in \mathcal{A}})^{\mathsf{T}} \left(\tilde{\mathbf{A}}^{\mathsf{T}}\right)^{+} \tilde{\pi}_{\odot}^{-1}$.

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

Algorithm 3: Identifies a mixture of product distributions given $3k-3$ ζ-separated observable bits.

- Line 5: This line is a critical place where it is necessary to acknowledge the empirical nature of the quantities we are computing with, and account for this explicitly. The matrix $\tilde{\mathbf{C}}_{\mathcal{BA}}$ will w.h.p. be close in Frobenius norm to $\mathbf{C}_{\mathcal{BA}}$ when the moments are computed from a large sample, but $\tilde{\mathbf{C}}_{\mathcal{BA}}$ will almost certainly have $\operatorname{rank}(\tilde{\mathbf{C}}_{\mathcal{BA}}) = 2^{k-1} > \operatorname{rank}(\mathbf{C}_{\mathcal{BA}}) = k$, which will introduce instability in the pseudoinverse. To avoid this we replace $\tilde{\mathbf{C}}_{\mathcal{BA}}$ (resp. $\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}$) with $\hat{\mathbf{C}}_{\mathcal{BA}}$ (resp. $\hat{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}$), the rank k approximation obtained by truncating to the first k singular values. The SVD of an $m \times m$ matrix can be computed in time $O(m^3)$ [48] which in our setting is $O(2^{3k})$. Note that computing the pseudoinverse of $\hat{\mathbf{C}}_{\mathcal{BA}}$ takes no more time than computing $\hat{\mathbf{C}}_{\mathcal{BA}}$ since we use the same SVD of $\tilde{\mathbf{C}}_{\mathcal{BA}}$ to compute both. (There are also methods which are faster for a low-rank approximation.)

- Line 6: If $\sigma_k(\mathbf{C}_{\mathcal{BA}})$ is too small, we have a contradiction to our assumption that S and T consisted of ζ-separated observables. Likewise, $\sigma_k(\mathbf{C}_{\mathcal{BA}})$ being too small contradicts the ζ-separation of the variables in S and T'.

- Lines 7-8: Here we create v_0 and v_1, both of which are directly constructed from observable quantities.

- Lines 9-15: In lines 9-10, we do the first step of the iteration that we continue in lines 11-15. We initialize u_1 and u_1' by multiplying v_1 by the pseudoinverses of $\mathbf{C}_{\mathcal{BA}}$ and $\mathbf{C}_{\mathcal{B}'\mathcal{A}}$, respectively. Note here a difference from [44], where the $\mathbf{C}_{\mathcal{BA}}$ matrices were $k \times k$ and full-rank. We then compute v_i, u_i, and u_i' for larger values of i. We use the fact that $v_{i+j} = (u_i \otimes u_j)\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$, proven in Lemma 69 to do this computation. We analyze the error introduced by this step in Lemma 80. This crucially relies on the condition number bound in Theorem 71, which is proved in Section 4.4. Each time we multiply by $\hat{\mathbf{C}}^{+}_{\mathcal{BA}}$ we introduce some additional error and we are only able to bound the error by showing that the operator norm of $\hat{\mathbf{C}}^{+}_{\mathcal{BA}} - \mathbf{C}^{+}_{\mathcal{BA}}$ is small, which requires this condition number bound. There are $\lg k$ steps in the iteration, each of which requires $2^{O(k)}$ time, since we multiply by matrices of size $2^{O(k)}$ constantly many times each iteration. (The Kronecker product computation also requires $2^{O(k)}$ time.)

- Lines 16-18: We assemble the power moments of X_1 into a Hankel matrix and invoke an algorithm for identifying a k-MixIID. (If the Hankel matrix has insufficiently large second smallest singular value, we conclude that one of our assumptions has not been satisfied—in particular, our estimated moments could be outside the required accuracy; or that there was a rare statistical error in the sampling.) The recovery of $\mathbf{m}_1$ and π can be done in time $k^{2+o(1)} + O(k(\log^2 k)\log\log(\varepsilon^{-1}))$ (see Corollary 83 for the runtime and the error bound).

- Lines 19-21: We compute $\mathbf{A}$ by multiplying the matrix of the first k v_i vectors by the inverse of the matrix $\mathrm{Vdm}(\tilde{\mathbf{m}}_1)$, the Vandermonde matrix generated from the empirical version of row $\mathbf{m}_1$. (See Lemma 86 for the resulting error bound.)

- Line 22: We compute $\mathbf{B}$ by multiplying $\mathbf{C}_{\mathcal{BA}}$ by the pseudoinverse of $\mathbf{A}$. Again, the pseudo-inverse here replaces the regular inverse used in [44] because $\mathbf{A}$ is no longer square. (See Lemma 88 for the resulting error bound.)

- Lines 23-24: We compute the remaining parameters in a similar fashion, using $\mathbf{A}^+$ or $\mathbf{B}^+$ as needed, depending on if the observable in question could overlap with an observable used by $\mathbf{A}$ or $\mathbf{B}$.

Lemma 69. *For all ℓ, $(v_\ell)_\emptyset = \mathbb{E}[X_1^{\odot\ell}]$.*

Proof. We show by induction on ℓ that $v_\ell = \mathbf{m}_1^{\odot\ell}\pi_\odot\mathbf{A}^\mathsf{T}$, hence $(v_\ell)_{\{1\}} = \mathbf{m}_1^{\odot\ell}\pi_\odot\mathbb{1}^\mathsf{T} = \mathbb{E}[X_1^\ell]$. The base case of $\ell = 1$ follows trivially from the definition of v_1. So suppose that this is true of $v_{\ell'}$ for $\ell' < \ell$. Let $\ell = 2^i + j$ for $j \in \{1, 2, \ldots, 2^i\}$. Now, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{B}'$ have full column rank, so

$$\begin{aligned}
v_\ell &= (v_j\mathbf{C}^+_{\mathcal{BA}} \otimes v_{2^i}\mathbf{C}^+_{\mathcal{B}'\mathcal{A}})\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} \\
&= (\mathbf{m}_1^{\odot j}\pi_\odot A^\mathsf{T}(\mathbf{B}\pi_\odot\mathbf{A}^\mathsf{T})^+ \otimes \mathbf{m}_1^{\odot 2^i}\pi_\odot A^\mathsf{T}(\mathbf{B}'\pi_\odot\mathbf{A}^\mathsf{T})^+)(\mathbf{B} * \mathbf{B}')\pi_\odot\mathbf{A}^\mathsf{T} \qquad \text{(using Defn. 61)} \\
&= (\mathbf{m}_1^{\odot j}(\mathbf{B})^+ \otimes \mathbf{m}_1^{\odot 2^i}(\mathbf{B}')^+)(\mathbf{B} * \mathbf{B}')\pi_\odot\mathbf{A}^\mathsf{T} \\
&= \mathbf{m}_1^\ell\pi_\odot\mathbf{A}^\mathsf{T},
\end{aligned}$$

□

When actually running the algorithm, we will only have access to the approximations $\tilde{\mathbf{C}}_{\mathcal{BA}}$, $\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}$, and $\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$, and, starting with $\tilde{v}_1$, we will compute

approximate vectors $\tilde{u}_j = \tilde{v}_j \hat{\mathbf{C}}^{+}_{\mathcal{B}\mathcal{A}}$, and $\tilde{u}'_{2^i} = \tilde{v}_{2^i} \hat{\mathbf{C}}^{+}_{\mathcal{B}'\mathcal{A}}$, and use those to compute an approximate vector $\tilde{v}_\ell = (\tilde{u}_j \otimes \tilde{u}'_{2^i}) \tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$. Notice the advantage of the recurrence in Line 12; we are able to get away with performing at most $1 + \lg k$ iterations to compute any of $\tilde{v}_1, \ldots, \tilde{v}_{2k}$. Each iteration consists of at most two matrix multiplications by $\hat{\mathbf{C}}^{+}_{\mathcal{B}\mathcal{A}}$ and $\hat{\mathbf{C}}^{+}_{\mathcal{B}'\mathcal{A}}$, followed by a convolution, followed by multiplication by $\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$. Each such step can increase the initial sampling error by at most a factor of $\zeta^{-O(k)} \pi_{\min}^{-O(1)}$. By starting with empirical moments accurate to within $\varepsilon \zeta^{\Omega(k \log k)} \pi_{\min}^{\Omega(\log k)}$, we can ensure that the resulting vectors $\tilde{v}_i$ are sufficiently close to the vectors v_i to start solving for $\mathbf{m}_1$ and π.

The conclusion of our analysis can be stated as follows:

Theorem 70. *1. If all multilinear moments used in the computation are within $\pm\varepsilon$ of their true values, the output $\tilde{\pi}$ and $\tilde{\mathbf{m}}$ will satisfy*

$$\|\tilde{\mathbf{m}} - \mathbf{m}\|_\infty, \|\tilde{\pi} - \pi\|_\infty \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

2. With high probability, all empirical multilinear moments appearing in Algorithm 3 will be within $\pm\varepsilon$ of their true value.

Proof. 1. This is an immediate consequence of Theorem 89 and Corollary 83.

2. With N samples, Hoeffding's inequality upper bounds the probability that each moment $g(S)$ differs from its empirical estimate $\tilde{g}(S)$ by at least ε as

$$\Pr\left[|g(\tilde{S}) - g(S)| > \varepsilon\right] \leq 2e^{-2\varepsilon^2 N}.$$

A union bound over $3(2^k)$ such moments gives us a bound for the failure event,

$$\Pr\left[\exists S \ : \ |g(\tilde{S}) - g(S)| > \varepsilon\right] \leq 6(2^k)e^{-2\varepsilon^2 N}.$$

This becomes constant when the number of samples N is $O(k/\varepsilon^2)$.

□

We note that a simpler version of the above procedure uses only $2k - 1$ ζ-separated variables, but needs $2k - 1$ iterations to compute the $\tilde{v}_i$-s, hence the required initial accuracy would be exponential in k^2 rather than in $k \log k$.

4.4 The Condition Number Bound

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

Theorem 71. *Let* $\mathbf{m}$ *have* $k-1$ *rows, every row* ζ*-separated, and let* $\mathbf{M} = \mathbb{H}(\mathbf{m})$. *List the singular values of* $\mathbf{M}\pi_\odot$ *in decreasing order:* $\sigma_1 \geq \sigma_2 \geq \ldots$, *and note* $\sigma_{k+1} = \ldots = \sigma_{2^{k-1}} = 0$. *Then*

$$\sigma_k \geq \frac{(\min \pi_i)(\zeta/2\sqrt{5})^{k-1}}{\sqrt{k(k^2+1)^{k-1}}}.$$

Definition 72. $\mathbb{H}_p = \{\mathbb{H}(\mathbf{n}) : \mathbf{n} \in \mathbb{R}^{[k-1]\times[p]}\}$. (So $\mathbb{H}_1$ consists of rank-1 tensors of order $k-1$.)

The proof of Theorem 71 relies on the following insight. Since $\mathbf{M}$ has dimensions $2^{k-1} \times k$, σ_k characterizes the least norm of $\mathbf{M}v$ ranging over any unit vector v, but it does *not* characterize the least norm of vectors of the form $h\mathbf{M}$; the left-kernel of $\mathbf{M}$ is of course very large. The insight is that it *does* become possible to bound σ_k in terms of such vectors h, *provided* h *is restricted to rank* 1 *tensors.*

Proof. Let $\tau(\mathbf{m}, \pi) = \min_{0\neq h\in\mathbb{H}_1} \left\|h^\mathsf{T}\mathbf{M}\pi_\odot\right\| / \|h\|$. The proof rests on two key lemmas:

Lemma 73. $\sigma_k(\mathbf{M}\pi_\odot) \geq \tau(\mathbf{m}, \pi)/\sqrt{k\prod_1^{k-1}(1+\|\mathbf{m}_{i*}\|^2)}$.

Proof. Consider $v \in \mathbb{R}^k$, $\|v\| = 1$, achieving σ_k, i.e., $r \coloneqq \mathbb{H}(\mathbf{m})\pi_\odot v$ satisfies $\|r\| = \sigma_k$. W.l.o.g. $\|v_k\| \geq 1/\sqrt{k}$. Now $\mathbb{H}(\mathbf{m})_{*k} = \frac{1}{\pi_k v_k}\left(r - \sum_{j=1}^{k-1}\pi_j v_j \mathbb{H}(\mathbf{m})_{*j}\right)$.

Now we carefully choose $h \in \mathbb{H}_1$ based on $\mathbf{m}$. Define $\mathbf{n} \in \mathbb{R}^{[k-1]}$ by $\mathbf{n}_i \coloneqq -1/\mathbf{m}_{ii}$; and let $h \coloneqq \left(\prod_1^{k-1}\mathbf{m}_{ii}\right)\mathbb{H}(\mathbf{n})$. For $j \neq k$ we have

$$\begin{aligned} h^\mathsf{T}\mathbb{H}(\mathbf{m})_{*j} &= \left(\prod_1^{k-1}\mathbf{m}_{ii}\right)\sum_S \mathbf{n}_S\mathbf{m}_{S,j} = \left(\prod_1^{k-1}\mathbf{m}_{ii}\right)\sum_S(-1)^{|S|}\left(\prod_{i\in S}1/\mathbf{m}_{ii}\right)\left(\prod_{i\in S}\mathbf{m}_{ij}\right) \\ &= \sum_S(-1)^{|S|}\left(\prod_{i\notin S}\mathbf{m}_{ii}\right)\left(\prod_{i\in S}\mathbf{m}_{ij}\right) = \prod_{i=1}^{k-1}(\mathbf{m}_{ii}-\mathbf{m}_{ij}) = 0. \end{aligned}$$

So, $h^\mathsf{T}\mathbb{H}(\mathbf{m})_{*j}\pi_j = 0$ for $j = 1, \ldots, k-1$. For $j = k$ we can also show that $h^\mathsf{T}\mathbb{H}(\mathbf{m})_{*j}\pi_j$ is small:

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

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

$$\begin{aligned}\left\|h^{\mathsf{T}}\mathbb{H}(\mathbf{m})\pi_{\odot}\right\| &= \left|(h^{\mathsf{T}}\mathbb{H}(\mathbf{m})\pi_{\odot})_k\right| = \frac{1}{v_k}h^{\mathsf{T}}r \leq \sqrt{k}\,\|h\|\,\|r\| \\ &= \sqrt{k}\,\|h\|\,\sigma_k = \sigma_k\sqrt{k\prod_1^{k-1}(1+\mathbf{m}_{ii}^2)} \quad \text{due to } h \text{ being a rank 1 tensor}\end{aligned}$$

□

Lemma 74. $\tau(\mathbf{m},\pi) \geq (\min \pi_i)(\zeta/2\sqrt{5})^{k-1}$.

Proof. Consider any $\mathbf{G} \in \mathbb{H}_1$, say $\mathbf{G} = \mathbb{H}(\mathbf{g})$, $\mathbf{g} \in \mathbb{R}^{[k-1]}$. Then $(\mathbf{G}^{\mathsf{T}}\mathbb{H}(\mathbf{m})\pi_{\odot})_j = \pi_j \sum_S \mathbf{g}_S\mathbf{m}_{S,j} = \pi_j \prod_1^{k-1}(1+\mathbf{g}_i\mathbf{m}_{i,j})$. 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. $\mathbf{m}_{ij} < \zeta$; exclude these j's. Next, for each i for which $\mathbf{g}_i < \frac{1}{2}$, there is at most one j s.t. $\left|\frac{1}{\mathbf{g}_i} + \mathbf{m}_{ij}\right| < \zeta/2$; 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 consider three cases for i. In each case we lower bound $\left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right|$.

1. $\mathbf{g}_i \geq 1/2, \mathbf{m}_{ij} \geq \zeta$. Then $\left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right| \geq \mathbf{m}_{ij} \geq \zeta$.
2. $-1/2 \leq \mathbf{g}_i < 1/2$. Then $\left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right| \geq \sqrt{\frac{(\mathbf{m}_{ij}-2)^2}{5}} \geq 1/\sqrt{5}$. (This does not depend on j.)
3. $\mathbf{g}_i < -1/2, |\frac{1}{\mathbf{g}_i} + \mathbf{m}_{ij}| \geq \zeta/2$. Then $\left|\frac{1+\mathbf{g}_i\mathbf{m}_{ij}}{\sqrt{1+\mathbf{g}_i^2}}\right| = \left|\frac{\mathbf{g}_i(\frac{1}{\mathbf{g}_i}+\mathbf{m}_{ij})}{\mathbf{g}_i\sqrt{1+1/\mathbf{g}_i^2}}\right| \geq \frac{\zeta}{2\sqrt{5}}$.

We therefore have $\tau(\mathbf{m},\pi) \geq (\min \pi_i)(\zeta/2\sqrt{5})^{k-1}$. □

The proof of Theorem 71 now follows from Lemmas 73, 74, and the fact that $\|\mathbf{m}_{i*}\| \leq k$ for any i. □

4.5 Analysis of the Algorithm

Here we will bound the errors introduced in each step of Algorithm 3, referring often to both the empirical and idealized versions of each quantity in question. First, we will introduce some notation to simplify some intermediate calculations:

Definition 75. $\zeta_1 := \zeta/9k^{3/2}$.

We will work under the following assumptions when doing the analysis.

Assumption 1. All empirical multilinear moments appearing in computations are within an additive ε of their true values.

Assumption 2. The sets S, T, T' contain ζ-separated observables.

Finally, for a matrix M, $\|M\|$ will denote the spectral norm. We will frequently use (implicitly) the upper-bound $\|M\| \leq \|M\|_F$, where $\|\cdot\|_F$ is the Frobenius norm.

Bounding the Operator Norm and Condition Number for the Ideal Matrices and Vectors

First, we argue that the ideal versions of the matrices under consideration are well-behaved, crucially relying on Theorem 71.

Lemma 76. *The matrices* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{B}'$ *satisfy*

1. $\mathbf{A}_{\emptyset *} = \mathbf{B}_{\emptyset *} = \mathbf{B}'_{\emptyset *} = \mathbb{1}$, *the all ones row vector.*
2. $\sigma_k(\mathbf{A}), \sigma_k(\mathbf{B}), \sigma_k(\mathbf{B}') \geq \zeta_1^k$.
3. $\sigma_{\max}(\mathbf{A}), \sigma_{\max}(\mathbf{B}), \sigma_{\max}(\mathbf{B}') \leq k2^{k-1}$.

Moreover, the matrices $\mathbf{C}_{\mathcal{BA}}$ *and* $\mathbf{C}_{\mathcal{B}'\mathcal{A}}$ *satisfy*

1. $\sigma_{\max}(\mathbf{C}_{\mathcal{BA}}), \sigma_{\max}(\mathbf{C}_{\mathcal{B}'\mathcal{A}}) \leq 2^{2k-2}$.
2. $\sigma_k(\mathbf{C}_{\mathcal{BA}}), \sigma_k(\mathbf{C}_{\mathcal{B}'\mathcal{A}}) \geq \pi_{\min}\zeta_1^{2k}$.

Proof. This follows immediately from Theorem 71, the definitions of $\mathbf{C}_{\mathcal{BA}}$ and $\mathbf{C}_{\mathcal{B}'\mathcal{A}}$, and the min-max characterization of the first and last singular values. □

Corollary 77. $\|(\mathbf{C}_{\mathcal{BA}})^+\|, \|(\mathbf{C}_{\mathcal{B}'\mathcal{A}})^+\| \leq \pi_{\min}^{-1}\zeta_1^{-2k}$.

Lemma 78. $\|u_i\|, \|u'_i\| \leq \pi_{\min}^{-1}\zeta_1^{-2k}$, *and* $\|v_i\| \leq \zeta_1^{-k}$.

Proof. Clearly, $\|v_i\| \leq 2^{k-1} \leq \zeta_1^{-k}$, as v_i is a vector of 2^{k-1} moments of products of Bernoulli random variables. Now $\|u_i\| = \left\|v_i\mathbf{C}^+_{\mathcal{BA}}\right\| \leq \zeta_1^{-k}\left\|\mathbf{C}^+_{\mathcal{BA}}\right\| \leq \pi_{\min}^{-1}\zeta_1^{-2k}$. A similar argument bounds $\|u'_i\|$. □

Bounding Error in Derived Quantitites

Lemma 79. *If ε is sufficiently small,*

$$\left\|\hat{\mathbf{C}}_{\mathcal{BA}} - \mathbf{C}_{\mathcal{BA}}\right\|_2, \left\|\hat{\mathbf{C}}_{\mathcal{B}'\mathcal{A}} - \mathbf{C}_{\mathcal{B}'\mathcal{A}}\right\|_2, \left\|\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} - \mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\|_2 \leq 2^{3k}\varepsilon < \zeta_1^{-k}\varepsilon,$$

and

$$\left\|\hat{\mathbf{C}}^+_{\mathcal{BA}} - \mathbf{C}^+_{\mathcal{BA}}\right\|_2, \left\|\hat{\mathbf{C}}^+_{\mathcal{B}'\mathcal{A}} - \mathbf{C}^+_{\mathcal{B}'\mathcal{A}}\right\|_2 \leq 2\pi_{\min}^{-2}\zeta_1^{-5k}\varepsilon.$$

Proof. To prove the first inequalities we use the fact that $\left\|\hat{\mathbf{C}}_{\mathcal{BA}} - \mathbf{C}_{\mathcal{BA}}\right\|_2 \leq 2\left\|\tilde{\mathbf{C}}_{\mathcal{BA}} - \mathbf{C}_{\mathcal{BA}}\right\|$ since $\left\|\hat{\mathbf{C}}_{\mathcal{BA}} - \tilde{\mathbf{C}}_{\mathcal{BA}}\right\| \leq \left\|\tilde{\mathbf{C}}_{\mathcal{BA}} - \mathbf{C}_{\mathcal{BA}}\right\|$ (as the nearest rank k matrix to $\tilde{\mathbf{C}}_{\mathcal{BA}}$ can be no further than the distance from $\tilde{\mathbf{C}}_{\mathcal{BA}}$ to $\mathbf{C}_{\mathcal{BA}}$). We then observe that $\|\cdot\|_2 \leq \|\cdot\|_F$ and the fact that every entry in the matrix of differences is at most ε in magnitude. For the final inequality we use Lemmas 90 and 91 to get

$$\left\|\hat{\mathbf{C}}^+_{\mathcal{BA}} - \mathbf{C}^+_{\mathcal{BA}}\right\| \leq 2\left\|\mathbf{C}^+_{\mathcal{BA}}\right\|^2\left\|\hat{\mathbf{C}}_{\mathcal{BA}} - \mathbf{C}_{\mathcal{BA}}\right\| \leq 2\pi_{\min}^{-2}\zeta_1^{-5k}\varepsilon.$$

The same argument bounds $\left\|\hat{\mathbf{C}}^+_{\mathcal{B}'\mathcal{A}} - \mathbf{C}^+_{\mathcal{B}'\mathcal{A}}\right\|_2$. □

Lemma 80. *For any $i \in \lg(2k)$ and $j \leq 2^i$,*

$$\|\tilde{v}_j - v_j\|, \|\tilde{u}_j - u_j\|, \left\|\tilde{u}'_j - u'_j\right\| \leq \pi_{\min}^{-2i}\zeta_1^{-11ik}\varepsilon.$$

Proof. Recall that we initialize the algorithm with

$$\tilde{v}_1 \leftarrow (\tilde{\mathrm{g}}(R \cup \{1\}))_{R\in\mathcal{A}}, \qquad \tilde{u}_1 \leftarrow \tilde{v}_1\hat{\mathbf{C}}^+_{\mathcal{BA}}, \qquad \tilde{u}'_1 \leftarrow \tilde{v}_1\hat{\mathbf{C}}^+_{\mathcal{BA}}.$$

First, we observe that $\|\tilde{v}_1 - v_1\| \leq \varepsilon$ by assumption. Since $\tilde{u}_1, \tilde{u}'_1$ are computed in the same manner here as in the loop, we will bound that error in the induction. Now assume that the claim holds up to $i-1$, and let ε_{i-1} be the bound obtained in the $i-1$st step. Recall that in each iteration of the outer loop we compute

$$\tilde{v}_{2^i} \leftarrow (\tilde{u}_{2^{i-1}} \otimes \tilde{u}'_{2^{i-1}})\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}, \qquad \tilde{u}_{2^i} \leftarrow \tilde{v}_{2^i}\hat{\mathbf{C}}^+_{\mathcal{BA}}, \qquad \tilde{u}'_{2^i} \leftarrow \tilde{v}_{2^i}\hat{\mathbf{C}}^+_{\mathcal{B}'\mathcal{A}}.$$

We also observe that we choose ε so that $\|\tilde{u}_j - u_j\| \ll \|u_j\|$ and $\|\tilde{v}_i - v_i\| \ll \|v_j\|$ for all i, j so that we can upper bound $\|\tilde{u}_j\| \leq 2\|u_j\| \leq 2\pi_{\min}^{-1}\zeta_1^{-2k}$ and $\|\tilde{v}_i\| \leq 2\|v_i\| \leq 2\zeta_1^{-k}$. We will first focus on bounding $\|\tilde{v}_{2^i} - v_{2^i}\|$. To do this we write

$$\tilde{v}_{2^i} - v_{2^i} = (\tilde{u}_{2^{i-1}} \otimes \tilde{u}'_{2^{i-1}})\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} - (u_{2^{i-1}} \otimes u'_{2^{i-1}})\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}.$$

We now let $w = \tilde{u}_{2^{i-1}} - u_{2^{i-1}}$, $w' = \tilde{u}'_{2^{i-1}} - u_{2^{i-1}}$, and $E = \tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} - \mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$, and use the triangle inequality to obtain

$$\|\tilde{v}_{2^i} - v_{2^i}\| \leq \left\|(w \otimes \tilde{u}'_{2^{i-1}})\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| + \left\|(\tilde{u}_{2^{i-1}} \otimes w')\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| + \|(\tilde{u}_{2^{i-1}} \otimes \tilde{u}'_{2^{i-1}})E\|.$$

The first two terms can each be bounded by

$$\|w\|\, 2\, \|u_1\| \left\|\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| \leq 2\pi_{\min}^{-1}\zeta_1^{-2k}2^{3k}\varepsilon_{i-1},$$

and the final term is bounded by

$$4\, \|u_1\|^2\, \|E\| \leq 4\pi_{\min}^{-2}\zeta_1^{-4k}\zeta_1^{-k}\varepsilon.$$

As a result we get that

$$\|\tilde{v}_{2^i} - v_{2^i}\| \leq 4\pi_{\min}^{-2}\zeta_1^{-5k}(\varepsilon_{i-1} + \varepsilon).$$

Now we can bound $\|\tilde{u}_{2^i} - u_{2^i}\|$ by observing that

$$\tilde{u}_{2^i} - u_{2^i} = \tilde{v}_{2^i}\hat{\mathbf{C}}^+_{\mathcal{BA}} - v_{2^i}\mathbf{C}^+_{\mathcal{BA}}.$$

Let $z = \tilde{v}_{2^i} - v_{2^i}$ and let $D = \hat{\mathbf{C}}^+_{\mathcal{BA}} - \mathbf{C}^+_{\mathcal{BA}}$. The above equation becomes

$$\tilde{u}_{2^i} - u_{2^i} = (v_{2^i} + z)(\mathbf{C}^+_{\mathcal{BA}} + D) - v_{2^i}\mathbf{C}^+_{\mathcal{BA}} = v_{2^i}D + z\mathbf{C}^+_{\mathcal{BA}} + zD,$$

and after taking norms and using the triangle inequality we obtain

$$\|\tilde{u}_{2^i} - u_{2^i}\| \leq \|v_{2^i}D\| + \left\|z\mathbf{C}^+_{\mathcal{BA}}\right\| + \|zD\|.$$

By Corollary 77, Lemma 78 and the induction hypothesis, we get

$$\begin{aligned}
\|\tilde{u}_{2^i} - u_{2^i}\| &\leq \pi_{\min}^{-2}\zeta_1^{-k}2\zeta_1^{-5k}\varepsilon + \|\tilde{v}_{2^i} - v_{2^i}\|\left(\pi_{\min}^{-1}\zeta_1^{-2k} + 2\pi_{\min}^{-2}\zeta_1^{-5k}\varepsilon\right) \\
&\leq 2\pi_{\min}^{-2}\zeta_1^{-6k}\varepsilon + 4\pi_{\min}^{-2}\zeta_1^{-5k}(\varepsilon_{i-1} + \varepsilon)(\pi_{\min}^{-1}\zeta_1^{-2k} + 2\pi_{\min}^{-2}\zeta_1^{-5k}\varepsilon) \\
&\leq 32\pi_{\min}^{-2}\zeta_1^{-10k}\varepsilon_{i-1} \\
&\leq \pi_{\min}^{-2}\zeta_1^{-11k}\varepsilon_{i-1},
\end{aligned}$$

where we use the fact that $\varepsilon_{i-1} \geq \varepsilon$. For j not a power of 2, we can do the same analysis, and since the error bound is increasing in j, the result will follow. □

Corollary 81. *Algorithm 3 will produce vectors $\tilde{v}_i$ for $i \leq 2k$ satisfying*

$$\|\tilde{v}_i - v_i\| \leq \pi_{\min}^{-O(\lg k)}\zeta^{-O(k\lg k)}\varepsilon.$$

Applying the k-MixIID Algorithm.
We use the following theorem.

Theorem 82 (A slight restatement of Theorem 22 from Chapter 2)**.** *Given a mixture $\mathcal{M} = (m, \pi)$ of k Bernoulli random variables with probabilities $m_1, \ldots, m_k$ and mixing probabilities $\pi_1, \ldots, \pi_k$, respectively, let $[\mathcal{H}_{k+1}]_{i,j=0}^{k} = \mu_{i+j}$ be the matrix of moments of the distribution. If m is ζ-separated, then there is an algorithm, LearnPowerDistribution, that takes as input a Hankel matrix $[\tilde{\mathcal{H}}_{k+1}]_{i,j=0}^{k} = \tilde{\mu}_{i+j}$ of approximate moments of $\mathcal{M}$ satisfying $\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_2 \leq \pi_{\min} 2^{-\gamma} \zeta^{16k}$ (for some $\gamma \geq 1$), and outputs a model $\tilde{\mathcal{M}} = (\tilde{m}, \tilde{\pi})$ satisfying*

$$\|\tilde{m} - m\|_\infty, \|\tilde{\pi} - \pi\|_\infty \leq 2^{-\gamma}$$

using $O(k^2 \log k + k \log^2 k \cdot \log(\log \zeta^{-1} + \log \pi_{\min}^{-1} + \gamma))$ arithmetic operations.

Corollary 83. *The output $(\tilde{\mathbf{m}}_1, \tilde{\pi})$ of LearnPowerDistribution in line 14 of Algorithm 3 satisfies*

$$\|\tilde{\mathbf{m}}_1 - \mathbf{m}_1\|, \|\tilde{\pi} - \pi\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

Proof. Every entry $(\tilde{v}_i)_{\{1\}}$ satisfies $\|(\tilde{v}_i)_1 - (v_i)_1\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon$ so

$$\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon$$

which implies that

$$\|\tilde{\mathbf{m}}_1 - \mathbf{m}_1\|_\infty, \|\tilde{\pi} - \pi\|_\infty \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

Finally, we need to multiply by $k \leq \zeta^{-1}$ to account for the conversion to the Euclidean norm to get the stated bound. □

Solving for the Rest of the Model.
Once we have computed $\tilde{\mathbf{m}}_1$ and $\tilde{\pi}$, we will use them to compute the remaining model parameters. In this section we bound the additional error introduced by these computations.

Proposition 84. $\|\mathrm{Vdm}(\tilde{\mathbf{m}}_1) - \mathrm{Vdm}(\mathbf{m}_1)\| \leq k \|\tilde{\mathbf{m}}_1 - \mathbf{m}_1\| \leq \zeta^{-1} \|\tilde{\mathbf{m}}_1 - \mathbf{m}_1\|$.

Claim 85 (Claim 26 in [47])**.** $\|\mathrm{Vdm}(\mathbf{m}_1)^{-1}\| \leq 2^k/\zeta^{k-1} \leq \zeta^{-2k}$ *when $\mathbf{m}_1$ is ζ-separated.*

Lemma 86. *The computed* $\tilde{\mathbf{A}}$ *produced by Algorithm 3 will satisfy*

$$\left\|\tilde{\mathbf{A}} - \mathbf{A}\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

Proof. Recall $\tilde{V} = (\tilde{v}_0; \ldots; \tilde{v}_{k-1})$ from Algorithm 3 and $V = (v_0; \ldots; v_{k-1})$ is its real-value analog. First, we observe that $\left\|\tilde{V}\right\| \leq \zeta^{-3k}$. Also, by Lemma 90 and Claim 85, $\|\mathrm{Vdm}(\tilde{\mathbf{m}}_1)^{-1}\| \leq \zeta^{-3k}$. Now, by Corollary 83, $\left\|\tilde{\pi}_{\odot}^{-1} - \pi_{\odot}^{-1}\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon$. Thus,

$$\left\|\tilde{\pi}_{\odot}^{-1} - \pi_{\odot}^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\right\| \left\|\tilde{V}\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

Now, by Lemma 91, $\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1} - (\mathrm{Vdm}(\mathbf{m}_1))^{-1}\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon$, so,

$$\left\|\tilde{\pi}_{\odot}^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1} - (\mathrm{Vdm}(\mathbf{m}_1))^{-1}\right\| \left\|\tilde{V}\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

Finally, $\left\|\tilde{V} - V\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon$, so that

$$\left\|\pi_{\odot}^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\right\| \left\|\tilde{V} - V\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

Putting these together, we easily obtain

$$\begin{aligned}
\left\|\tilde{\mathbf{A}} - \mathbf{A}\right\| &= \left\|\tilde{\pi}_{\odot}^{-1}(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\tilde{V} - \pi_{\odot}^{-1}(\mathrm{Vdm}(\mathbf{m}_1))^{-1}V\right\| \\
&\leq \left\|\tilde{\pi}_{\odot}^{-1} - \pi_{\odot}^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\right\| \left\|\tilde{V}\right\| + \left\|\tilde{\pi}_{\odot}^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1} - (\mathrm{Vdm}(\mathbf{m}_1))^{-1}\right\| \left\|\tilde{V}\right\| \\
&\quad + \left\|\pi_{\odot}^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\right\| \left\|\tilde{V} - V\right\| \left(\left\|\mathrm{Vdm}(\mathbf{m}_1)^{-1}E_2\right\| + \|E_1 V\|_{\infty}\right) \\
&\quad + \|w\| \left\|\mathrm{Vdm}(\tilde{\mathbf{m}}_1)^{-1}\right\| \left\|\tilde{V}\right\| \\
&\leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.
\end{aligned}$$

□

Lemma 87. $\left\|(\tilde{\mathbf{A}}^{\mathsf{T}})^{+}\right\| \leq \zeta^{-O(k)}$ *and* $\left\|(\tilde{\mathbf{B}}^{\mathsf{T}})^{+}\right\| \leq \zeta^{-O(k)}$.

Proof. We prove this for $(\tilde{\mathbf{A}}^{\mathsf{T}})^{+}$, the argument for $(\tilde{\mathbf{B}}^{\mathsf{T}})^{+}$ is the same.

Notice that Lemma 86 together with the lower bound on $\sigma_k(\mathbf{A})$ from Theorem 71 imply that $\tilde{\mathbf{A}}^{\mathsf{T}}$ has full row rank. Now,

$$\left\|(\tilde{\mathbf{A}}^{\mathsf{T}})^{+}\right\| \leq \left\|(\mathbf{A}^{\mathsf{T}})^{+}\right\| + \left\|(\tilde{\mathbf{A}}^{\mathsf{T}})^{+} - (\mathbf{A}^{\mathsf{T}})^{+}\right\|.$$

We bound each term separately. We have

$$\left\|(\mathbf{A}^\mathsf{T})^+\right\| = \left\|\mathbf{A}(\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}\right\| \leq \|\mathbf{A}\| \left\|(\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}\right\| \leq k2^{k-1}/(\sigma_k(\mathbf{A}))^2 \leq \zeta^{-O(k)}.$$

As for the second term, notice that

$$\begin{aligned}\left\|\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T} - \mathbf{A}\mathbf{A}^\mathsf{T}\right\| &= \left\|\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^\mathsf{T} - \mathbf{A}^\mathsf{T}) + (\tilde{\mathbf{A}} - \mathbf{A})\mathbf{A}^\mathsf{T}\right\| \\ &\leq \left\|\tilde{\mathbf{A}}\right\| \left\|\tilde{\mathbf{A}}^\mathsf{T} - \mathbf{A}^\mathsf{T}\right\| + \left\|\tilde{\mathbf{A}} - \mathbf{A}\right\| \|\mathbf{A}^\mathsf{T}\| \\ &\leq \pi_{\min}^{-O(\lg k)}\zeta^{-O(k\lg k)}\varepsilon.\end{aligned}$$

Therefore, by Lemma 91 also $\left\|(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T})^{-1} - (\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}\right\| \leq \pi_{\min}^{-O(\lg k)}\zeta^{-O(k\lg k)}\varepsilon$. Thus, we get for the second term

$$\begin{aligned}\left\|(\tilde{\mathbf{A}}^\mathsf{T})^+ - (\mathbf{A}^\mathsf{T})^+\right\| &= \left\|\tilde{\mathbf{A}}(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T})^{-1} - \mathbf{A}(\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}\right\| \\ &= \left\|(\tilde{\mathbf{A}} - \mathbf{A})(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T})^{-1} + \mathbf{A}((\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T})^{-1} - (\mathbf{A}\mathbf{A}^\mathsf{T})^{-1})\right\| \\ &\leq \left\|\tilde{\mathbf{A}} - \mathbf{A}\right\| \left\|(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T})^{-1}\right\| + \|\mathbf{A}\| \left\|(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T})^{-1} - (\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}\right\| \\ &\leq \pi_{\min}^{-O(\lg k)}\zeta^{-O(k\lg k)}\varepsilon.\end{aligned}$$

□

Lemma 88. $\left\|\tilde{\mathbf{B}} - \mathbf{B}\right\| \leq \pi_{\min}^{-O(\lg k)}\zeta^{-O(k\lg k)}\varepsilon.$

Proof. We can bound $\tilde{\mathbf{B}} - \mathbf{B}$ using the same tools as in the previous bounds. First, we bound

$$\left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}} - \mathbf{C}_{\mathcal{B}\mathcal{A}}\right\| \left\|(\tilde{\mathbf{A}}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1}\right\| \leq \pi_{\min}^{-1}\zeta^{-O(k)}\varepsilon.$$

Next,

$$\left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}\right\| \left\|(\tilde{\mathbf{A}}^\mathsf{T})^+ - (\mathbf{A}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1}\right\| \leq \pi_{\min}^{-O(\lg k)}\zeta^{-O(k\lg k)}\varepsilon.$$

Finally,

$$\left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}\right\| \left\|(\tilde{\mathbf{A}}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1} - \pi_\odot^{-1}\right\| \leq \pi_{\min}^{-O(\lg k)}\zeta^{-O(k\lg k)}\varepsilon.$$

The resulting bound is

$$\begin{aligned}\left\|\tilde{\mathbf{B}} - \mathbf{B}\right\| &= \left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}(\tilde{\mathbf{A}}^\mathsf{T})^+\tilde{\pi}_\odot^{-1} - \mathbf{C}_{\mathcal{B}\mathcal{A}}(\mathbf{A}^\mathsf{T})^+\pi_\odot\right\| \\ &\leq \left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}} - \mathbf{C}_{\mathcal{B}\mathcal{A}}\right\| \left\|(\tilde{\mathbf{A}}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1}\right\| + \\ &\quad + \left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}\right\| \left\|(\tilde{\mathbf{A}}^\mathsf{T})^+ - (\mathbf{A}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1}\right\| + \\ &\quad + \left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}\right\| \left\|(\tilde{\mathbf{A}}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1} - \pi_\odot^{-1}\right\| \\ &\leq \pi_{\min}^{-O(\lg k)}\zeta^{-O(k\lg k)}\varepsilon.\end{aligned}$$

□

Theorem 89. *Algorithm 3 will compute $\tilde{\mathbf{m}}_i$ satisfying, for all $i \in [n]$,*

$$\|\tilde{\mathbf{m}}_i - \mathbf{m}_i\|_\infty \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

Proof. We will compute the bound using the inversion of $\tilde{\mathbf{B}}^\mathsf{T}$ since this will give us the worst case. Let $\tilde{y} = (\tilde{\mathrm{g}}(R \cup \{i\}))_{R \in \mathcal{B}}$, and let $y = (\mathrm{g}(R \cup \{i\}))_{R \in \mathcal{B}}$. We note that $\|\tilde{y} - y\| \leq \zeta^{-O(k)} \varepsilon$, by assumption on the sample size, and $\|\tilde{y}\| \leq 2^{k-1}$. Then,

$$\|\tilde{y} - y\| \left\|(\tilde{\mathbf{B}}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1}\right\| \leq \pi_{\min}^{-O(1)} \zeta^{-O(k)} \varepsilon,$$

$$\|\tilde{y}\| \left\|(\tilde{\mathbf{B}}^\mathsf{T})^+ - (\mathbf{B}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1}\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon,$$

and

$$\|\tilde{y}\| \left\|(\tilde{\mathbf{B}}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1} - \pi_\odot^{-1}\right\| \leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.$$

So that we get

$$\begin{aligned}
\|\tilde{\mathbf{m}}_i - \mathbf{m}_i\| &= \left\|\tilde{y}(\tilde{\mathbf{B}}^\mathsf{T})^+ \tilde{\pi}_\odot^{-1} - y(\mathbf{B}^\mathsf{T})^+ \pi_\odot^{-1}\right\| \\
&\leq \|\tilde{y} - y\| \left\|(\tilde{\mathbf{B}}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1}\right\| + \|\tilde{y}\| \left\|(\tilde{\mathbf{B}}^\mathsf{T})^+ - (\mathbf{B}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1}\right\| + \|\tilde{y}\| \left\|(\tilde{\mathbf{B}}^\mathsf{T})^+\right\| \left\|\tilde{\pi}_\odot^{-1} - \pi_\odot^{-1}\right\| \\
&\leq \pi_{\min}^{-O(\lg k)} \zeta^{-O(k \lg k)} \varepsilon.
\end{aligned}$$

□

Miscellaneous Proofs

Lemma 90. *Let $M, \tilde{M} \in \mathbb{R}^{n \times n}$ have rank $k \leq n$. Let $\left\|\tilde{M} - M\right\| = \varepsilon \leq \sigma_k(M)/2$. Then $\left\|\tilde{M}^+\right\| \leq 2 \left\|M^+\right\|$.*

Proof. We observe that

$$\left\|\tilde{M}^+\right\| = \frac{1}{\sigma_k(\tilde{M})} \leq \frac{1}{\sigma_k(M) - \sigma_k(M)/2} \leq 2 \left\|M^+\right\|.$$

□

Lemma 91 (Theorem 3.4 in [49])**.** *Let $M, \tilde{M} \in \mathbb{R}^{n \times n}$ have rank $k \leq n$. Then $\left\|M^+ - \tilde{M}^+\right\| \leq 2 \left\|M^+\right\| \left\|\tilde{M}^+\right\| \left\|M - \tilde{M}\right\|$.*

Chapter 5

THE IDENTIFIABILITY OF UNIFORM MIXTURES OF BINOMIAL DISTRIBUTIONS WITH LOG-LINEAR INFLUENCES

In the work thus far we have been exclusively concerned with the setting where there is a single latent variable U which induces independence among $X_1, \ldots, X_n$ when it is conditioned upon. In general Bayesian networks we are unlikely to have a single predecessor for most nodes and so this setting is insufficiently general to handle many cases of interest. However, this setting does allow us to identify the parameters of the joint distribution on U and the X_i using only $O(k)$ moments, when U is supported on k states.

Towards the goal of eventually understanding identifiability in more general graphs, we will return the k-Mix IID setting and start generalizing from there.

We begin by introducing multiple latent predecessors $U_1, \ldots, U_\ell$; we will have each visible variable X be conditionally independent from the others given the values of each of the U_i. With ℓ hidden variables having k states each, we then obtain a setting in which there k^ℓ possible assignments to the latent variables, and each such assignment induces a (potentially different) probability that $X = 1$. One method for identifying parameters in this case (excluding the measure zero set of parameters for which the probability that $X = 1$ is the same across two or more assignments to $U_1, \ldots, U_\ell$) is to run a k^ℓ-Mix IID algorithm. Unfortunately, doing this requires observing $O(k^\ell)$ copies of X and requires learning k^ℓ parameters. Here, there is no advantage to be gained by knowing that there are multiple independent latent variables, as opposed to a single latent variable with k^ℓ states. We will call this the naïve setting.

We would like to understand when we can hope to take advantage of the independence structure among the latent variables, which requires that there be some observable trace of this indpendence structure in the Bernoulli parameters induced by the assignments to the latent variables. Towards that end, we will consider the case where the contributions to the Bernoulli parameters induced by an assignment can be decoupled into a contribution

for each latent variable. The simplest way to do this is to have $\Pr(X = 1 \mid u_1, \ldots, u_\ell) = \prod_{i=1}^{\ell} \alpha_{i,u_i}$ for parameters $\{\alpha_{i,j}\}_{i,j \in [\ell] \times [k]}$ so that the contributions from each latent variable are multiplied together; changing one latent variable only changes that term in the product. This model is already interestingly different from the naïve setting when $k = 2$. Then we have $2\ell + \ell$ parameters describing the joint distribution in this setting, compared to $2^\ell + \ell$ parameters in the naïve setting. We can hope to do better than just running Prony's algorithm with a mixture of 2^ℓ distributions in this case.

This assumption of "independent influence" of the latent variables on the observed variables makes the resulting model into a log-linear model; we have

$$\log \Pr(X = 1 \mid u_1, \ldots, u_\ell) = \sum_{i=1}^{\ell} \log \alpha_{i,u_i}.$$

While we have a log-linear dependence of the probability on the parameters, these are unnormalized, in that the probability of $X = 1 \mid u_1, \ldots, u_\ell$ does not depend on the sum of the probabilities of $X = 1$ across all assignments to $U_1, \ldots, U_\ell$. As a result, these are in fact distinct from many of the models that go under the heading of "log-linear models", such as restricted Boltzmann machines or Markov random fields.

Formalizing our model We are concerned with the following mixture model: We have n binary random variables $X_1, \ldots, X_n$ that are iid conditioned on a collection of ℓ independent latent binary variables $U_1, \ldots, U_\ell$. Each U_i is uniform over $\{0, 1\}$.[1] Finally, we have that the the probability that $X = 1$ given $U = u$ is given by the product of terms corresponding to each of the hidden variables, namely $P(X = x \mid U = u) = \prod_{j=1}^{\ell} \alpha_{j,u_j}$. We have 2ℓ parameters α_{ib} for $i = 1, 2, \ldots, \ell$ and $b = 0, 1$.

The probability that $X = 1$ is given by $2^{-\ell}$ times

$$\mu_1 \coloneqq \prod_{i=1}^{\ell} (\alpha_{i0} + \alpha_{i1}).$$

With enough copies of X we have access to

$$\mu_j \coloneqq \prod_{i=1}^{\ell} (\alpha_{i0}^j + \alpha_{i1}^j)$$

[1] Relaxing this assumption makes the analysis quite a bit more difficult, and we do not know if the same result will hold in that setting.

for arbitrary j.[2]

We are going to throw out the restrictions on the statistical model, and consider the problem of solving the polynomial system.

We first observe that μ_1 is invariant to the transformation

$$(\alpha_{i0}, \alpha_{i1}) \mapsto (\lambda\alpha_{i0}, \lambda\alpha_{i1}),$$
$$(\alpha_{j0}, \alpha_{j1}) \mapsto (\lambda^{-1}\alpha_{i0}, \lambda^{-1}\alpha_{i1})$$

for any $\lambda > 0$ and $i \neq j$ (where all other parameters are unchanged). As a result, we can without loss of generality solve for parameters satisfying $\alpha_{i0} + \alpha_{i1} = \mu_1^{\frac{1}{\ell}}$ for all $i \in [\ell]$. Let $\gamma \coloneqq \mu_1^{\frac{1}{\ell}}$.

To solve for α_{i0}, α_{i1} it now suffices to solve for $a_i \coloneqq \alpha_{i0}\alpha_{i1}$, the product of our two parameters. In particular, we get that $\alpha_{i0}(\gamma - \alpha_{i0}) = a_i$ which we can easily solve for α_{i0}.

We will obtain the following theorem:

Theorem 92. *The mapping from $\{a_i\}$ to $\mu_2, \ldots, \mu_{\ell+1}$ is identifiable.*

Our immediate goal will now be to understand the polynomial system of equations in the variables $\{a_i\}$ that we obtain after this transformation.

First, we observe that

$$\alpha_{i0}^2 + \alpha_{i1}^2 = (\alpha_{i0} + \alpha_{i1})^2 - 2\alpha_{i0}\alpha_{i1} = \gamma^2 - 2a_i$$

so that

$$\mu_2 = \prod_{i=1}^{\ell}(\gamma^2 - 2a_i).$$

Moreover,

$$\begin{aligned}\alpha_{i0}^3 + \alpha_{i1}^3 &= (\alpha_{i0} + \alpha_{i1})^3 - 3(\alpha_{i0}^2\alpha_{i1} + \alpha_{i0}\alpha_{i1}^2) \\ &= \gamma^3 - 3\alpha_{i0}\alpha_{i1}(\alpha_{i0} + \alpha_{i1}) \\ &= \gamma^3 - 3\gamma a_i.\end{aligned}$$

In subsequent computations, I'll omit the subscript 'i' everywhere since it remain constant throughout.

[2]The actual observation we can make is of $\mu_j/2^\ell$, which we can multiply by 2^ℓ to get μ_j.

Definition 93. Let $p_m(a)$ denote the univariate polynomial in a obtained by simplifying $\alpha_0^n + \alpha_1^n$ using $\alpha_0 + \alpha_1 = \gamma$ and $a = \alpha_0\alpha_1$.

Proposition 94.

$$p_0 = 2, \tag{5.1}$$

$$p_1 = \gamma, \tag{5.2}$$

$$p_m = \gamma p_{m-1} - a p_{m-2}, \quad m \geq 2. \tag{5.3}$$

Proof. First, $p_0 = \alpha_0^0 + \alpha_1^0 = 2$ and $p_1 = \alpha_0^1 + \alpha_1^1 = \gamma$. Now

$$\begin{aligned} p_m &= \alpha_0^m + \alpha_1^m \\ &= (\alpha_0^{m-1} + \alpha_1^{m-1})(\alpha_0 + \alpha_1) - \alpha_0\alpha_1(\alpha_0^{m-2} + \alpha_1^{m-2}) \\ &= \gamma p_{m-1} - a p_{m-2}. \end{aligned}$$

□

Observation 95. The polynomials defined above satisfy the following:

1. The degree of p_m is $\lfloor m/2 \rfloor$.
2. $p_m(0) = \gamma^m$ for $m > 0$.
3. The leading term of p_m has a negative sign if $\lfloor m/2 \rfloor$ is odd, and a positive sign if $\lfloor m/2 \rfloor$ is even.

Proof. By induction over m. □

Proposition 96. *For any $m > 2$,*

1. *the roots $r_1, \ldots, r_{\lfloor m/2 \rfloor}$ of p_m are all real, distinct, and contained in $(0, \infty)$.*
2. *Let $s_1, \ldots, s_{\lfloor (m-1)/2 \rfloor}$ be the roots of p_{m-1}. Then $0 < r_1 < s_1$ and $s_{i-1} < r_i < s_i$ for $i = 2, \ldots, \lfloor m/2 \rfloor$ and $s_{\lfloor (m-1)/2 \rfloor} < r_{\lfloor m/2 \rfloor}$ when p_m has degree greater than p_{m-1}.*

Proof. We will induct over m. The claim clearly holds for $m = 0, 1$. Now $p_2 = -2a + \gamma^2$ which has a root at $\gamma^2/2$ and $p_3 = \gamma(-2a + \gamma^2) - a\gamma = -2\gamma a + \gamma^3 - \gamma a = -3\gamma a + \gamma^3$ which has a root at $\gamma^2/3$.

Now fix an arbitrary $m \in \mathbb{N}$. Let $d \coloneqq \deg(p_{m-2})$, so $d+1 = \deg(p_m)$. Let $s_1, \ldots, s_d$ be the roots of p_{m-2} and let $r_1, \ldots, r_d$ be the first d roots of p_{m-1}. By the inductive hypothesis, $0 < r_1 < s_1 < r_2 < \cdots < r_d < s_d$. If $\deg(p_{m-1}) = d+1 > \deg(p_m)$, then there is an additional root r_{d+1} of p_{m-1} with $s_d < r_{d+1}$. Since we have accounted for every root of p_{m-1} and p_{m-2}, it must be the case that the value of p_{m-1} alternates between strictly positive and strictly negative on the sequence of open intervals $(-\infty, r_1), (r_1, r_2), (r_2, r_3), \ldots, (r_{\lfloor (m-1)/2 \rfloor}, \infty)$ and p_{m-2} alternates in sign on the intervals $(-\infty, s_1), (s_1, s_2), \ldots, (s_d, \infty)$. Now we compute

$$\begin{aligned} p_m(r_1) &= \gamma p_{m-1}(r_1) - r_1 p_{m-2}(r_1) &&= -r_1 p_{m-2}(r_1) < 0; \\ p_m(s_1) &= \gamma p_{m-1}(s_1) - s_1 p_{m-2}(s_1) &&= \gamma p_{m-1}(s_1) < 0. \end{aligned}$$

By the preceding observation, $p_m(0) = \gamma^m > 0$ so there must be a root of p_m in $(0, r_1)$.

For $1 < i < d$, $r_i \in (s_{i-1}, s_i)$ and $s_i \in (r_i, r_{i+1})$. Moreover, $p_m(r_i) = -r_i p_{m-2}(r_i)$ and $p_m(s_i) = \gamma p_{m-1}(s_i)$, so $\operatorname{sign}(p_m(r_i)) = \operatorname{sign}(p_m(s_i)) = -\operatorname{sign}(p_m(r_{i+1}))$. We conclude that there is a root of p_m in the interval (s_i, r_{i+1}) for $1 < i < d$.

we have thus shown that there are roots of p_m in each of the intervals

$$(0, r_1), (s_1, r_2), (s_2, r_3), \ldots, (s_{d-1}, r_d).$$

If $\deg(p_{m-1}) = d+1$, then we also get that there is a root in (s_d, r_{d+1}) and we are done. If $\deg(p_{m-1}) = d$, then the leading term of p_m has a different sign than the leading terms of p_{m-1} and p_{m-2}. Since $\operatorname{sign}(p_m(s_d)) = \operatorname{sign}(p_{m-1}(s_d))$ and s_d is greater than all the roots of p_{m-1} it must be the case that $\operatorname{sign}(p_{m-1}(x)) = \operatorname{sign}(p_{m-1}(s_d))$ for all $x \in [s_d, \infty)$. But $\lim_{x\to\infty} p_m(x) = -\lim_{x\to\infty} p_{m-1}(x)$, so there must be a root of p_m in (s_d, ∞). we have thus accounted for all $d+1$ roots of p_m, proving the claim. □

The identifiability of symmetric systems of polynomial equations

Consider a sequence of ℓ univariate polynomials $p_1, \ldots, p_\ell$. Let $x_1, \ldots, x_\ell$ be indeterminates.

We can construct symmetric polynomials in the x_i by taking products as follows:

$$q_i(x) \coloneqq \prod_{j=1}^{\ell} p_i(x_j).$$

Proposition 97. *Suppose that for all $i \in [\ell]$, p_i has a smallest positive real root α_i with multiplicity 1 that is not a root of $p_1, p_2, \ldots, p_{i-1}$. Then the mapping*

$$(y_1, \ldots, y_\ell) \mapsto (q_1(y), \ldots, q_\ell(y))$$

is invertible, except on a set of measure zero.

Proof. In what follows, let $\#(\alpha_j, p_i)$ denote the multiplicity of root α_j in p_i. By assumption, $\#(\alpha_j, p_i) = 0$ for $j > i$ and $\#(\alpha_j, p_j) = 1$ for all j.

We will now construct a sequence of rational functions $r_1, \ldots, r_\ell$ satisfying

$$\#(\alpha_j, r_i) = \mathbb{1}_{i=j},$$

where $\#(\alpha_j, r_i) < 0$ if α_j is a pole of r_i.

First, we set $r_1 \coloneqq p_1$, since $\#(\alpha_j, p_1) = 0$ for all $j > 1$.

Now inductively we construct r_i as follows:

$$r_i \coloneqq p_i \left(\prod_{i'=1}^{i-1} r_{i'}^{-\#(\alpha_{i'}, p_i)} \right).$$

By construction, $\#(\alpha_j, r_i) = 0$ for $j < i$ and $\#(\alpha_i, r_i) = 1$. Moreover, $\#(\alpha_j, r_i) = 0$ for $j > i$ since

$$\#(\alpha_j, p_i) = \#(\alpha_j, r_1) = \cdots = \#(\alpha_j, r_{i-1}) = 0.$$

Now define $s_i \coloneqq \prod_{j=1}^{\ell} r_i(y_j)$ for $i = 1, \ldots, \ell$ so that s_i is the symmetric product of r_i evaluated at each indeterminate, just as q_i is the symmetric product of p_i evaluated at each indeterminate.

In fact, we have

$$s_i = q_i \left(\prod_{i'=1}^{i-1} s_{i'}^{-\#(\alpha_{i'}, p_i)} \right).$$

Let F and G be the following mappings:

$$(y_1, \ldots, y_\ell) \overset{F}{\mapsto} (q_1, \ldots, q_\ell) \overset{G}{\mapsto} (s_1, \ldots, s_\ell).$$

Finally, let $H = G \circ F$.

We now consider the Jacobian of H, evaluated at the point $\alpha = (\alpha_1, \ldots, \alpha_\ell)$.

By construction we have that

$$\frac{\partial s_i}{\partial y_j}(\alpha) = \left(\prod_{j' \neq j} r_i(\alpha_{j'})\right) r_i'(\alpha_j).$$

Now

$$\prod_{j' \neq j} r_i(\alpha_{j'}) \neq 0 \iff i = j$$

since $r_i(\alpha_i) = 0$ and $\#(\alpha_j, r_i) = 0$ for any $j \neq i$. Moreover, $r_i'(\alpha_i) \neq 0$ since α_i is a simple root of r_i, so we conclude that $\frac{\partial s_i}{\partial y_j}(\alpha) \neq 0 \iff i = j$. Thus, the Jacobian is a diagonal matrix with non-zero diagonal entries and is thus invertible. We conclude that there is an open neighborhood around α in which H is invertible, which implies the same for F. Since F is a polynomial map, we conclude it is generically identifiable. □

Proof of Theorem 92. Finally, Theorem 92 follows immediately from Propositions 94 and 97. □

Conjecture 98. *Let $p_1, \ldots, p_\ell$ be univariate polynomials and let $q_i := \prod_{j=1}^{\ell} p_i(x_j)$ for $i = 1, \ldots, \ell$. Let $\alpha_1, \ldots, \alpha_L$ be the set of all roots appearing in any p_i. Consider the matrix $M \in \mathbb{R}^{\ell \times L}$ with entry $M_{ij} \in \mathbb{Z}_{\geq 0}$ being the multiplicity with which α_j appears as a root in p_i. Then the mapping $(x_1, \ldots, x_\ell) \mapsto (q_1(x), \ldots, q_\ell(x))$ is invertible if and only if M has rank ℓ over $\mathbb{Q}$.*

BIBLIOGRAPHY

[1] F. B. Hildebrand, *Introduction to Numerical Analysis*, 2nd. McGraw-Hill, 1974.

[2] B. Simon, *A Comprehensive Course in Analysis*. American Mathematical Society, 2015.

[3] 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 *Int'l Conference on Research in Computational Molecular Biology*, ser. Lecture Notes in Computer Science, L. Cowen, Ed., vol. 11457, Springer, 2019, pp. 136–157. DOI: 10.1007/978-3-030-17083-7_9.

[4] A. Feldmann and W. Whitt, "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models," *Performance Evaluation*, vol. 31, no. 3, pp. 245–279, 1998.

[5] J. Pearl, *Causality*, 2nd. Cambridge, 2009.

[6] A. Anandkumar, D. J. Hsu, and S. M. Kakade, "A method of moments for mixture models and hidden Markov models," in *Proceedings of the 25th Annual Conference on Learning Theory - COLT*, ser. JMLR Proceedings, vol. 23, 2012, pp. 33.1–33.34. [Online]. Available: http://proceedings.mlr.press/v23/anandkumar12/anandkumar12.pdf.

[7] R. de Prony, "Essai expérimentale et analytique," *J. Écol. Polytech.*, vol. 1, no. 2, pp. 24–76, 1795.

[8] R. Kumaresan, D. W. Tufts, and L. L. Scharf, "A Prony method for noisy data: Choosing the signal components and selecting the order in exponential signal models," *Proceedings of the IEEE*, vol. 72, no. 2, pp. 230–233, 1984.

[9] Y. Rabani, L. J. Schulman, and C. Swamy, "Learning mixtures of arbitrary distributions over large discrete domains," in *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, 2014, pp. 207–224. DOI: 10.1145/2554797.2554818.

[10] J. Li, Y. Rabani, L. J. Schulman, and C. Swamy, "Learning arbitrary statistical mixtures of discrete distributions," in *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, 2015, pp. 743–752. DOI: 10.1145/2746539.2746584.

[11] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 289–296.

[12] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, no. 2, pp. 217–235, 2000.

[13] S. Arora, R. Ge, and A. Moitra, "Learning topic models — going beyond SVD," in *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, 2012.

[14] A. Anandkumar, D. P. Foster, D. J. Hsu, S. M. Kakade, and Y.-K. Liu, "A spectral algorithm for latent Dirichlet allocation," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 917–925.

[15] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie, "On the learnability of discrete distributions," in *Proceedings of the 26th Annual ACM Symposium on Theory of Computing*, 1994, pp. 273–282. DOI: 10.1145/195058.195155.

[16] Y. Freund and Y. Mansour, "Estimating a mixture of two product distributions," in *Proceedings of the 12th Annual Conference on Computational Learning Theory*, Jul. 1999, pp. 53–62. DOI: 10.1145/307400.307412.

[17] 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.*, vol. 31, no. 2, pp. 375–397, 2001. DOI: 10.1137/S0097539798342496.

[18] K. Chaudhuri and S. Rao, "Learning mixtures of product distributions using correlations and independence," in *Proceedings of the 21st Annual Conference on Learning Theory - COLT*, Omnipress, 2008, pp. 9–20. [Online]. Available: http://colt2008.cs.helsinki.fi/papers/7-Chaudhuri.pdf.

[19] J. Feldman, R. O'Donnell, and R. A. Servedio, "Learning mixtures of product distributions over discrete domains," *SIAM J. Comput.*, vol. 37, no. 5, pp. 1536–1564, 2008. DOI: 10.1137/060670705.

[20] S. Chen and A. Moitra, "Beyond the low-degree algorithm: Mixtures of subcubes and their applications," in *Proceedings of the 51st Annual ACM Symposium on Theory of Computing*, 2019, pp. 869–880.

[21] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference*. MIT Press, 2017.

[22] T. S. Chihara, *An Introduction to Orthogonal Polynomials*. Gordon and Breach, 1978.

[23] K. Schmüdgen, *The Moment Problem* (Graduate Texts in Mathematics). Springer International Publishing, 2017, vol. 277.

[24] Wikipedia contributors, *Hoeffding's inequality — Wikipedia, the free encyclopedia*, [Online; accessed 6-April-2020], 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Hoeffding%27s_inequality&oldid=946643344.

[25] T. J. Rivlin, "Bounds on a polynomial," *Journal of Research of the National Bureau of Standards - B. Math. Sci.*, vol. 74B, no. 1, pp. 47–54, Jan. 1970.

[26] F. D. Parker, "Inverses of Vandermonde matrices," *The American Mathematical Monthly*, vol. 71, no. 4, pp. 410–411, 1964.

[27] J. H. Wilkinson, "The perfidious polynomial," in *Studies in Numerical Analysis*, ser. Studies in Mathematics, G. H. Golub, Ed., vol. 24, Mathematical Association of America, 1984, pp. 1–28.

[28] V. Y. Pan and Z. Q. Chen, "The complexity of the matrix eigenproblem," in *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, 1999, pp. 507–516.

[29] W. Gautschi, "How (un)stable are Vandermonde systems," *Asymptotic and computational analysis*, vol. 124, pp. 193–210, 1990.

[30] S. G. Bartels and D. J. Higham, "The structured sensitivity of Vandermonde-like systems," *Numerische Mathematik*, vol. 62, pp. 17–33, 1992.

[31] V. Y. Pan, Z. Q. Chen, and A. Zheng, "The complexity of the algebraic eigenproblem," Math. Science Research Institute, Berkeley California, Tech. Rep. 1998-071, 1998.

[32] B. N. Parlett, *The Symmetric Eigenvalue Problem*. Prentice-Hall, Inc., 1998.

[33] V. Y. Pan, "Optimal and nearly optimal algorithms for approximating polynomial zeros," *Computers & Mathematics with Applications*, vol. 31, no. 12, pp. 97–138, 1996.

[34] S. L. Gordon, B. Mazaheri, Y. Rabani, and L. J. Schulman, "Source identification for mixtures of product distributions," in *Proceedings of the 34th Annual Conference on Learning Theory - COLT*, ser. Proceedings of Machine Learning Research, vol. 134, PMLR, 2021, pp. 2193–2216. [Online]. Available: http://proceedings.mlr.press/v134/gordon21a.html.

[35] S. Newcomb, "A generalized theory of the combination of observations so as to obtain the best result," *American Journal of Mathematics*, vol. 8, no. 4, pp. 343–366, 1886.

[36] K. Pearson, "Contributions to the mathematical theory of evolution III," *Philosophical Transactions of the Royal Society of London (A.)*, vol. 185, pp. 71–110, 1894.

[37] B. S. Everitt and D. J. Hand, "Mixtures of discrete distributions," in *Finite Mixture Distributions*, Dordrecht: Springer Netherlands, 1981, pp. 89–105.

[38] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, Inc., 1985.

[39] B. G. Lindsay, *Mixture Models: Theory, Geometry and Applications*. 1995, pp. i–163.

[40] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annual Review of Statistics and Its Application*, vol. 6, no. 1, pp. 355–378, 2019. DOI: 10.1146/annurev-statistics-031017-100325.

[41] 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," Proceedings of the ISIT 2018 pp. 1051-1055), 2018. [Online]. Available: https://arxiv.org/abs/1807.05444.

[42] 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," Proceedings of the ISIT'18 pp. 1051-1055.), 2018. [Online]. Available: https://arxiv.org/abs/1807.05444.

[43] S. Chen and A. Moitra, "Beyond the low-degree algorithm: Mixtures of subcubes and their applications," in *Proceedings of the 51st Annual ACM Symposium on Theory of Computing*, 2019, pp. 869–880. DOI: 10.1145/3313276.3316375.

[44] S. L. Gordon, B. Mazaheri, Y. Rabani, and L. J. Schulman, "Source identification for mixtures of product distributions," in *Proceedings of the 34th Annual Conference on Learning Theory - COLT*, ser. Proceedings of the Machine Learning Research, vol. 134, PMLR, 2021, pp. 2193–2216. [Online]. Available: http://proceedings.mlr.press/v134/gordon21a.html.

[45] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden Markov models," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1460–1480, Sep. 2012. DOI: 10.1016/j.jcss.2011.12.025.

[46] S. L. Gordon and L. J. Schulman, "Hadamard extensions and the identification of mixtures of product distributions," *IEEE Transactions on Information Theory*, vol. 68, no. 6, pp. 4085–4089, 2022. DOI: 10.1109/TIT.2022.3146630.

[47] S. L. Gordon, B. Mazaheri, Y. Rabani, and L. J. Schulman, "The sparse Hausdorff moment problem, with application to topic models," 2020. [Online]. Available: https://arxiv.org/abs/2007.08101.

[48] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 4th. The Johns Hopkins University Press, 2013.

[49] G. W. Stewart, "On the perturbation of pseudo-inverses, projections and linear least squares problems," *SIAM Review*, vol. 19, no. 4, pp. 634–662, 1977. DOI: 10.1137/1019104.