Learning Mixtures of Arbitrary Distributions over Large Discrete Domains

Yuval Rabani
The Hebrew University
Jerusalem 91904, Israel.
yrabani@cs.huji.ac.il

Leonard J. Schulman*
Caltech
Pasadena, CA 91125, USA.
schulman@caltech.edu

Chaitanya Swamy†
Univ. Waterloo
Waterloo, ON N2L 3G1, Canada.
cswamy@math.uwaterloo.ca

ABSTRACT

We give an algorithm for learning a mixture of *unstructured* distributions. This problem arises in various unsupervised learning scenarios, for example in learning *topic models* from a corpus of documents spanning several topics. We show how to learn the constituents of a mixture of k arbitrary distributions over a large discrete domain $[n] = \{1, 2, \ldots, n\}$ and the mixture weights, using $O(n \operatorname{polylog} n)$ samples. (In the topic-model learning setting, the mixture constituents correspond to the topic distributions.)

This task is information-theoretically impossible for $k > 1$ under the usual sampling process from a mixture distribution. However, there are situations (such as the above-mentioned topic model case) in which each sample point consists of several observations from the same mixture constituent. This number of observations, which we call the *"sampling aperture"*, is a crucial parameter of the problem.

We obtain the *first* bounds for this mixture-learning problem *without imposing any assumptions on the mixture constituents.* We show that efficient learning is possible exactly at the information-theoretically least-possible aperture of $2k - 1$. Thus, we achieve near-optimal dependence on n and optimal aperture. While the sample-size required by our algorithm depends exponentially on k, we prove that such a dependence is *unavoidable* when one considers general mixtures.

A sequence of tools contribute to the algorithm, such as concentration results for random matrices, dimension reduction, moment estimations, and sensitivity analysis.

*Supported in part by NSF CCF-1038578, NSF CCF-0515342, NSA H98230-06-1-0074, and NSF ITR CCR-0326554.

†Supported in part by NSERC grant 32760-06, an NSERC Discovery Accelerator Supplement Award, and an Ontario Early Researcher Award.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ITCS'14, January 12–14, 2014, Princeton, New Jersey, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2698-8/14/01 ...$15.00.
http://dx.doi.org/10.1145/2554797.2554818.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems; G.2 [**Discrete Mathematics**]; G.3 [**Probability and Statistics**]

Keywords

Randomized algorithms; mixture learning; topic models; spectral techniques; moment methods; linear programming; convex geometry

1. INTRODUCTION

We give an algorithm for learning a mixture of *unstructured* distributions. More specifically, we consider the problem of learning a mixture of k arbitrary distributions over a large finite domain $[n] = \{1, 2, \ldots, n\}$. This finds applications in various unsupervised learning scenarios including *collaborative filtering* [29], and learning *topic models* from a corpus of documents spanning several topics [39, 11], which is often used as the prototypical motivating example for this problem. Our goal is to learn the probabilistic model that is hypothesized to generate the observed data. In particular, we learn the constituents of the mixture and their weights in the mixture. (In the topic models application, the mixture constituents are the topic distributions.)

It is information-theoretically impossible to reconstruct the mixture model from single-snapshot samples. Thus, our work relies on multi-snapshot samples. To illustrate, in the (pure documents) topic model introduced in [39], each document consists of a *bag of words* generated by selecting a topic with probability proportional to its mixture weight and then taking independent samples from this topic's distribution (over words); so n is the size of the vocabulary and k is the number of topics. Notice that typically n will be quite large, and substantially larger than k. Also, clearly, if very long documents are available, the problem becomes easy, as each document already provides a very good sample for the distribution of its topic. Thus, it is desirable to keep the dependence of the sample size on n as low as possible, while at the same time minimize what we call the *aperture*, which is the number of snapshots per sample point (i.e., words per document). These parameters govern both the applicability of an algorithm and its computational complexity.

Our results. We provide the *first* bounds for the mixture-learning problem *without making any limiting assumptions* on the mixture constituents. Let probability distributions $p^1, \ldots, p^k \in \Delta^{n-1}$ denote the k-mixture constituents, where Δ^{n-1} is the $(n-1)$-simplex, and $w_1, \ldots, w_k$ denote the mix-

ture weights. Our algorithm uses

$$O\left(\frac{k^3 n \ln n}{\epsilon^6}\right) + O\left(\frac{k^2 n \ln^6 n \ln\left(\frac{k}{\epsilon}\right)}{\epsilon^4}\right) + O\left(\frac{k}{\epsilon}\right)^{O(k^2)} \quad (1)$$

documents (i.e., samples) and reconstructs with high probability (see Theorem 4.1) each mixture constituent up to ℓ_1-error ϵ, and each mixture weight up to additive error ϵ. We make no assumptions on the constituents. The asymptotic notation hides factors that are polynomial in $w_{\min} := \min_t w_t$ and the *"width"* of the mixture (which intuitively measures the minimum variation distance between any two constituents). The three terms in (1) correspond to the requirements for the number of 1-, 2-, and $(2k-1)$-snapshots respectively. So we need aperture $2k-1$ only for a small part of the sample (and this is necessary).

Notably, we achieve *near-optimal dependence on n and optimal aperture*. To see this, and put our bounds in perspective, notice importantly that we recover the mixture constituents within ℓ_1*-distance* ϵ. One needs $\Omega\big(n/\epsilon^2\big)$ samples to learn even a single arbitrary distribution over $[n]$ (i.e., $k=1$) within ℓ_1-error ϵ; for larger k but fixed aperture (independent of n), a sample size of $\Omega(n)$ is *necessary* to recover even the expectation of the mixture distribution with constant ℓ_1-error. On the other hand, aperture $\Omega\big((n+k^2)\log nk\big)$ is sufficient for algorithmically trivial recovery of the model with constant ℓ_∞ error using few samples. Restricting the aperture to $2k-2$ makes recovery *impossible* to arbitrary accuracy (without additional assumptions): we show that there are two far-apart k-mixtures that generate exactly the same aperture-$(2k-2)$ sample distribution; moreover, we prove that with $O(k)$ aperture, an exponential in k sample size is *necessary* for arbitrary-accuracy reconstruction. These lower bounds hold even for $n=2$, and hence apply to arbitrary mixtures even if we allow $O(k\log n)$ aperture. Also, they apply even if we only want to construct a k-mixture source that is close in transportation distance to the true k-mixture source (as opposed to recovering the parameters of the true mixture). Section 6 presents these lower bounds. (Interestingly, an exponential in k sample-size lower bound is also known for the problem of learning a mixture of k Gaussians [36], but this lower bound applies for the parameter-recovery problem and not for reconstructing a mixture that is close to the true Gaussian mixture.)

Our work yields new insights into the mixture-learning problem that nicely complements the recent interesting work of [4, 3, 2]. These papers posit certain assumptions on the mixture constituents, use constant aperture, and obtain incomparable sample-size bounds: they recover the constituents up to ℓ_2 or ℓ_∞ error using sample size that is poly(k) and sublinear in (or independent of) n. An important new insight revealed by our work is that such bounds of constant aperture and poly(k) sample size are *impossible* to achieve for arbitrary mixtures. Moreover, if we seek to achieve ℓ_1-error ϵ, there are inputs for which their sample size is $\Omega(n^3)$ (or worse, again ignoring dependence on $w_{\min}$ and "width"; see Appendix B). This is a significantly poorer dependence on n compared to our near-linear dependence (so our bounds are better when n is large but k is small). To appreciate a key distinction between our work and [4, 3, 2], observe that with $\Omega(n^3)$ samples, the entire distribution on 3-snapshots can be estimated fairly accurately; the challenge in [4, 3, 2] is therefore to recover the model from this relatively noiseless data. In contrast, a major challenge for achieving ℓ_1-reconstruction with $O(n \operatorname{polylog} n)$ samples is to ensure that the error remains bounded despite the presence of very noisy data due to the small sample size, and we develop suitable machinery to achieve this.

We now give a rough sketch of our algorithm (see Section 3) and the ideas behind its analysis (Section 4). Let $P = (p^1, \ldots, p^k)$, $r = \sum_t w_t p^t$ be the expectation of the mixture, and $k' = \operatorname{rank}(p^1 - r, \ldots, p^k - r)$. We first argue that it suffices to focus on isotropic mixtures (Lemma 3.3). Our algorithm reduces the problem to the problem of learning *one-dimensional mixtures*. Note that this is a special case of the general learning problem that we need to be able to solve (since we do not make any assumptions about the rank of P). We choose k' random lines that are close to the affine hull, $\operatorname{aff}(P)$, of P and "project" the mixture on to these k' lines. We learn each projected mixture, which is a one-dimensional mixture-learning problem, and combine the inferred projections on these k' lines to obtain k points that are close to $\operatorname{aff}(P)$. Finally, we project these k' points on to Δ^{n-1} to obtain k distributions over $[n]$, which we argue are close (in ℓ_1-distance) to $p^1, \ldots, p^k$.

Various difficulties arise in implementing this plan. We first learn a good approximation to $\operatorname{aff}(P)$ using spectral techniques and 2-snapshots. We use ideas similar to [35, 6, 34], but our challenge is to show that the covariance matrix $A = \sum_t w_t (p^t - r)(p^t - r)^\dagger$ can be well-approximated by the empirical covariance matrix with only $O(n \ln^6 n)$ 2-snapshots. A random orthonormal basis of the learned affine space supplies the k' lines on which we project our mixture. Of course, we do not know P, so "projecting" on to a basis vector b actually means that we project snapshots from P on to b by mapping item i to b_i. For this to be meaningful, we need to ensure that if the mixture constituents are far apart in variation distance then their projections $(b^\dagger p^t)_{t\in[k]}$ are also well separated relative to the spread of the support $\{b_1, \ldots b_n\}$ of the one-dimensional distribution. In general, we can only claim a relative separation of $\Theta\big(\frac{1}{\sqrt{n}}\big)$ (since $\min_{t\neq t'} \|p^t - p^{t'}\|_2$ may be $\Theta\big(\frac{1}{\sqrt{n}}\big)$). We avoid this via a careful balancing act: we prove (Lemma 4.3) that the ℓ_∞ norm of unit vectors in $\operatorname{aff}(P)$ is $O\big(\frac{1}{\sqrt{n}}\big)$, and argue that this isotropy property suffices since b is close to $\operatorname{aff}(P)$.

Finally, a key ingredient of our algorithm (see Section 5) is to show how to solve the one-dimensional mixture-learning problem and learn the real projections $(b^\dagger p^t)_{t\in[k]}$ from the projected snapshots. This is technically the most difficult step and the one that requires aperture $2k-1$ (the smallest aperture at which this is possible). We show that the projected snapshots on b yield empirical moments of a related distribution and use this to learn the projections and the mixture weights via a method of moments (see, e.g., [25, 24, 31, 10, 36, 3]). One technical difficulty is that variation distance in Δ^{n-1} translates to transportation distance [42] in the one-dimensional projection. We use a combination of convex programming and numerical-analysis techniques to learn the projections from the empirical "directional" moments. In the process, we establish some novel properties about the *moment curve*—an object that plays a central role in convex and polyhedral geometry [8]—that may be of independent interest.

Related work. The past decade has witnessed tremendous progress in the theory of learning statistical mixture

models. The most striking example is that of learning mixtures of high dimensional Gaussians. Starting with Dasgupta's groundbreaking paper [20], a long sequence of improvements [21, 5, 41, 32, 1, 24, 13] culminated in the recent results [31, 10, 36] that essentially resolve the problem in its general form. In this vein, other highly structured mixture models, such as mixtures of discrete product distributions [33, 26, 18, 25, 14, 16] and similar models [18, 9, 37, 32, 19, 15, 22], have been studied intensively. One important difference between this line of work and ours is that the structure of those mixtures enables learning using single-snapshot samples, whereas this is impossible in our case. Another interesting difference between our setting and the work on structured models (and this is typical of most results on PAC-style learning) is that the amount of information in each sample point is roughly in the same ballpark as the information needed to describe the model. In our setting, the amount of information in each sample point is exponentially sparser than the information needed to describe the model to good accuracy. Thus, the topic-model learning problem motivates the natural question of inference from sparse samples. This issue is also encountered in collaborative filtering; see [34] for some related theoretical problems.

Recently and independently, [4, 3, 2] have considered much the same question as ours.[1] They make certain assumptions about the mixture constituents which makes it possible to learn the mixture with constant aperture and $\mathrm{poly}(n, k)$ sample size (for ℓ_1-error). In comparison with our work, their sample bounds are attractive in terms of k but come at the expense of added assumptions (which are necessary), and have a worse dependence on n.

The assumptions in [4, 3, 2] impose some limitations on the applicability of their algorithms. To understand this, it is illuminating to consider the case where all the p^ts lie on a line-segment in Δ^{n-1} as an illustration. This poses no problems for our algorithm: we recover the p^ts along with their mixture weights. However, as we show below, the algorithms in [4, 3, 2] all fail to reconstruct this mixture. Anandkumar et al. [3] solve the same problem that we consider, under the assumption that P (viewed as an $n \times k$ matrix) has rank k. This is clearly violated here, rendering their algorithm inapplicable. The other two papers [4, 2] consider the setting where each multi-snapshot is generated from a combination of mixture constituents [39, 28]: first a convex combination $\lambda \in \Delta^{k-1}$ is sampled from a mixture distribution $\mathcal{T}$ on Δ^{k-1}, then the snapshot is generated by sampling from the distribution $\sum_{t=1}^{k} \lambda_t p^t$. The goal is to learn the mixture constituents and the mixture distribution. (The problem we consider is the special case where $\mathcal{T}$ places weight w^t on the t-th vertex of Δ^{k-1}.) [4] posits a *ρ-separability* assumption on the mixture constituents, wherein each p^t has a unique "anchor word" i such that $p_i^t \geq \rho$ and $p_i^{t'} = 0$ for every $t' \neq t$, whereas [2] weakens this to the requirement that P has rank k. Both papers handle the case where $\mathcal{T}$ is the Dirichlet distribution (which gives the latent Dirichlet model [12]); [4] obtains results for other mixture distributions as well.

In order to apply these algorithms, we can view the input as being specified by two constituents, x and y, which are the end points of the line segment; $\mathcal{T}$ then places weight w_t on the convex combination $(\lambda_t, 1-\lambda_t)^\dagger$, where $p^t = \lambda_t x + (1-\lambda_t)y$. This $\mathcal{T}$ is far from the Dirichlet distribution, so [3] does not apply here. Suppose that x and y satisfy the ρ-separability condition. (Note that ρ may only be $O(\frac{1}{n})$, even if x and y have *disjoint* supports.) We can then apply the algorithm of Arora et al. [4]. But this *does not* recover $\mathcal{T}$; it returns the "topic correlation" matrix $\mathrm{E}_{\mathcal{T}}[\lambda\lambda^\dagger]$, which does not reconstruct the mixture (w, P).

This limitation should not be surprising since [4] uses constant aperture. Indeed, [4] notes that it is impossible to reconstruct $\mathcal{T}$ with arbitrary accuracy (with any constant aperture) even if one knows the constituents x and y. In this context, we remark that our earlier work [40] uses the approach presented in this paper and solves the problem for *arbitrary* mixtures of two distributions, yielding a crisp statement about the tradeoff between the sampling aperture and the accuracy with which $\mathcal{T}$ can be learnt.

Our methods bear some resemblance with the recent independent work of Gravin et al. [27] who consider the problem of recovering the vertices of a polytope from its directional moments. [27] solves this problem for a polynomial density function assuming that exact directional moments are available; they do not perform any sensitivity analysis for measuring the error in their output if one has noisy information. In contrast, we solve this problem given only noisy empirical moment statistics and using much smaller aperture, albeit when the polytope is a subset of the $(n-1)$-simplex and the distribution is concentrated on its vertices.

Finally, it is also pertinent to compare our mixture-learning problem with the problem of learning a mixture of product distributions (e.g., [25]). Multi-snapshot samples can be thought of as single-snapshot samples from the power distribution on $[n]^K$, where K is the aperture. The product distribution literature typically deals with samples spaces that are the product of many small cardinality components, whereas our problem deals with samples spaces that are the product of few large cardinality components.

2. PRELIMINARIES

2.1 Mixture sources, snapshots, and projections

Let $[n]$ denote $\{1, 2, \ldots, n\}$, and Δ^{n-1} denote the $(n-1)$-simplex $\{x \in \mathbb{R}^n_{\geq 0} : \sum_i x_i = 1\}$. A *$k$-mixture source* (w, P) *on* $[n]$ consists of k mixture constituents $P = (p^1, \ldots, p^k)$, where p^t has support $[n]$ for all $t \in [k]$, along with the corresponding mixture weights $w = (w_1, \ldots, w_k) \in \Delta^{k-1}$. An m-snapshot from (w, P) is obtained by choosing $t \in [k]$ according to the distribution w, and then choosing $i \in [n]$ m times independently according to the distribution p^t. The probability distribution on m-snapshots is thus a mixture of k power distributions on the product space $[n]^m$. We also consider mixture sources whose constituents are distributions on $\mathbb{R}$. A *k-mixture source* (w, P) *on* $\mathbb{R}$ consists of k mixture constituents $P = (p^1, p^2, \ldots, p^k)$, where each p^t is a probability distribution on $\mathbb{R}$, along with corresponding mixture weights $w = (w_1, \ldots, w_k) \in \Delta^{k-1}$.

Given a distribution p on $[n]$ and a vector $x \in \mathbb{R}^n$, we define the projection of p on x, denoted $\pi_x(p)$, to be the discrete distribution on $\mathbb{R}$ that assigns probability mass $\sum_{i:x_i=\beta} p_i$ to $\beta \in \mathbb{R}$. (Thus, $\pi_x(p)$ has support $\{x_1, \ldots, x_n\}$ and $\mathrm{E}[\pi_x(p)] = x^\dagger p$.) Given a k-mixture source (w, P) on $[n]$, we define the

[1] An earlier stage of this work, including the case $k = 2$ as well as some other results that are not subsumed by this paper, dates to 2007. The last version of that phase has been posted since May 2008 at [40]. The extension to arbitrary k is from 2012.

projected k-mixture source $(w, \pi_x(P))$ on $\mathbb{R}$ to be the k-mixture source on $\mathbb{R}$ given by $\big(w, (\pi_x(p^1), \ldots, \pi_x(p^k))\big)$.

We also denote by $(w, \mathrm{E}[\pi_x(P)])$ the distribution that assigns probability mass w_t to $\mathrm{E}[\pi_x(p^t)] = x^\dagger p^t$ for all $t \in [k]$. This is an example of what we call a *k-spike distribution*, which is a distribution on $\mathbb{R}$ that assigns positive probability mass to k points in $\mathbb{R}$.

2.2 Transportation distance for mixtures

Let $\big(w, (p^1, \ldots, p^k)\big)$ and $\big(\tilde{w}, (\tilde{p}^1, \ldots, \tilde{p}^\ell)\big)$ be k- and ℓ-mixture sources on $[n]$ respectively. The *transportation distance* (with respect to the total variation distance $\frac{1}{2}\|x-y\|_1$ on measures on Δ^{n-1}) between these two sources, denoted by $\mathrm{Tran}(w, P; \tilde{w}, \tilde{P})$, is the optimum value of the following linear program (LP).

$$\text{minimize} \quad \sum_{i=1}^{k}\sum_{j=1}^{\ell} x_{ij} \cdot \frac{1}{2}\|p^i - \tilde{p}^j\|_1 \qquad \text{subject to}$$

$$\sum_{j=1}^{\ell} x_{ij} = w_i \quad \forall i \in [k], \qquad \sum_{i=1}^{k} x_{ij} = \tilde{w}_j \quad \forall j \in [\ell], \qquad x \geq 0.$$

The transportation distance $\mathrm{Tran}(w, \alpha; \tilde{w}, \tilde{\alpha})$ between a k-spike distribution $\big(w, \alpha = (\alpha_1, \ldots, \alpha_k)\big)$ and an ℓ-spike distribution $\big(\tilde{w}, \tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_\ell)\big)$ is defined as the optimum value of the above LP with the objective function replaced by $\sum_{i\in[k], j\in[\ell]} x_{ij}|\alpha_i - \tilde{\alpha}_j|$.

2.3 Perturbation results and operator norm of random matrices

Definition 1. The *operator norm* of A (induced by the ℓ_2 norm) is defined by $\|A\|_{\mathrm{op}} = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}$. The *Frobenius norm* of $A = (A_{i,j})$ is defined by $\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$.

LEMMA 2.1 (Weyl; see Theorem 4.3.1 in [30]). *Let A and B be $n \times n$ matrices such that $\|A - B\|_{\mathrm{op}} \leq \rho$. Let $\lambda_1(A) \geq \ldots \geq \lambda_n(A)$, and $\lambda_1(B) \geq \ldots \geq \lambda_n(B)$ be the sorted list of eigenvalues of A and B respectively. Then $|\lambda_i(A) - \lambda_i(B)| \leq \rho$ for all $i = 1, \ldots, n$.*

LEMMA 2.2. *Let A, B be $n\times n$ positive semi-definite (PSD) matrices whose nonzero eigenvalues are at least $\varepsilon > 0$. Let Π_A and Π_B be the projection operators onto the column spaces of A and B respectively. Let $\|A - B\|_{\mathrm{op}} \leq \rho$. Then $\|\Pi_A - \Pi_B\|_{\mathrm{op}} \leq \sqrt{4\rho/\varepsilon}$.*

PROOF. Note that $A\Pi_A = A$, $\Pi_A^2 = \Pi_A$, $B\Pi_B = B$, and $\Pi_B^2 = \Pi_B$. Let x be a unit vector. Since $\|(A-B)\|_{\mathrm{op}} \leq \rho$ and Π_B is a contraction, $\|(A - B)\Pi_B x\| \leq \rho\|\Pi_B x\| \leq \rho$. Now note that $(A - B)\Pi_B x = A\Pi_B x - Bx$ so by the triangle inequality, we have $\|A\Pi_B x - Ax\| \leq 2\rho$. Now we can also write $A\Pi_B x - Ax = A(\Pi_B - I)x = A(\Pi_A\Pi_B - \Pi_A)x$. Since A here is acting on a vector that has already been projected down by Π_A, we can conclude

$$2\rho \geq \|A\Pi_B x - Ax\| = \|A(\Pi_A\Pi_B - \Pi_A)x\| \geq \varepsilon\|(\Pi_A\Pi_B - \Pi_A)x\|.$$

Thus, $2\rho/\varepsilon \geq \|(\Pi_A - \Pi_A\Pi_B)x\|$. By the symmetric argument we also can write $2\rho/\varepsilon \geq \|(\Pi_B - \Pi_B\Pi_A)x\|$. Adding these and applying the triangle inequality we have

$$\begin{aligned} 4\rho/\varepsilon &\geq \|(\Pi_A - \Pi_A\Pi_B + \Pi_B - \Pi_B\Pi_A)x\| \\ &= \|(\Pi_A^2 - \Pi_A\Pi_B - \Pi_B\Pi_A + \Pi_B^2)x\| \\ &= \|(\Pi_A - \Pi_B)^2 x\| \qquad \square \end{aligned}$$

THEOREM 2.3 ([43]). *For every $\mu > 0$, there is a constant $\kappa = \kappa(\mu) = O(\mu) > 0$ such that the following holds. Let $X_{i,j}, 1 \leq i \leq j \leq n$ be independent random variables with $|X_{ij}| \leq K$, $\mathrm{E}[X_{i,j}] = 0$, and $\mathrm{Var}(X_{i,j}) \leq \sigma^2$ for all $i, j \in [n]$, where $\sigma \geq \kappa^2 n^{-1/2} K \ln^2 n$. Let A be the symmetric matrix with entries $A_{i,j} = X_{\min(i,j),\max(i,j)}$ for all $i, j \in [n]$. Then, $\Pr\big[\|A\|_{\mathrm{op}} \leq 2\sigma\sqrt{n} + \kappa(K\sigma)^{1/2} n^{1/4} \ln n\big] \geq 1 - n^{-\mu}$.*

3. OUR ALGORITHM

We now describe our algorithm that uses 1-, 2-, and $(2k-1)$-snapshots from the mixture source (w, P). Given a matrix Z, we use $\mathrm{Span}(Z)$ to denote the column space of Z. Let $r = \sum_{t=1}^{k} w_t p^t$ denote the 1-snapshot distribution of (w, P). Let M be the $n \times n$ symmetric matrix representing the 2-snapshot distribution of (w, P); so $M_{i,j}$ is the probability of obtaining the 2-snapshot $(i, j) \in [n]^2$. Let $R = rr^\dagger$.

PROPOSITION 3.1. *$M = \sum_{t=1}^{k} w_t p^t p^{t\dagger} = R + A$, where $A = \sum_{t=1}^{k} w_t (p^t - r)(p^t - r)^\dagger$.*

Note that M and A are both PSD. We say that (w, P) is *ζ-wide* if (i) $\|p - q\|_2 \geq \frac{\zeta}{\sqrt{n}}$ for any two distinct $p, q \in P$; and (ii) the smallest non-zero eigenvalue of A is at least $\zeta^2\|r\|_\infty \geq \frac{\zeta^2}{n}$. We assume that $w_{\min} := \min_t w_t > 0$. Let $k' = \mathrm{rank}(A) \leq k-1$. It is easy to estimate r using Chernoff bounds (Lemma A.1).

LEMMA 3.2. *For every $\mu \in \mathbb{N}$ and every $\sigma > 0$, if we use $N \geq \frac{8(\mu+2)}{\sigma^3} \cdot n \ln n$ independent 1-snapshots and set $\tilde{r}_i$ to be the frequency of i in these 1-snapshots for all $i \in [n]$, then with probability at least $1 - n^{-\mu}$ the following hold.*

$$\begin{aligned} (1-\sigma)r_i \leq \tilde{r}_i \leq (1+\sigma)r_i \qquad & \forall i \text{ with } r_i \geq \frac{\sigma}{2n} \\ \tilde{r}_i \leq (1+\sigma)\sigma/2n \qquad & \forall i \text{ with } r_i < \frac{\sigma}{2n}. \end{aligned} \tag{2}$$

It will be convenient in the sequel to assume that our mixture source (w, P) is *isotropic*, by which we mean that $\frac{1}{2n} \leq r_i \leq \frac{2}{n}$ for all $i \in [n]$; notice that this implies that $p_i^t \leq \frac{2}{w_{\min} n}$ for all $i \in [n]$. We show below that this can be assumed at the expense of a small additive error.

LEMMA 3.3. *Suppose that we can learn, with probability $1 - \frac{1}{\omega}$, the constituents of an isotropic ζ-wide k-mixture source on $[n]$ to within transportation distance ϵ using $N_1(n; \zeta, \omega, \epsilon)$, $N_2(n; \zeta, \omega, \epsilon)$, and $N_{2k-1}(n; \zeta, \omega, \epsilon)$ 1-, 2-, and $(2k-1)$-snapshots respectively. Then, we can learn, with probability $1 - O\big(\frac{1}{\omega}\big)$, the constituents of an arbitrary ζ-wide k-mixture source (w, P) on $[n]$ to within transportation distance 2ϵ using $O\big(\frac{\ln\omega}{\sigma^3} \cdot n \ln n\big) + 6\omega N_1\big(\frac{n}{\sigma}, \frac{\zeta}{2}, \omega, \epsilon\big)$, $6\omega N_2\big(\frac{n}{\sigma}, \frac{\zeta}{2}, \omega, \epsilon\big)$, and $6\omega N_{2k-1}\big(\frac{n}{\sigma}, \frac{\zeta}{2}, \omega, \epsilon\big)$ 1-, 2-, and $(2k-1)$-snapshots respectively, where $\sigma = \frac{\epsilon\zeta^2}{32k w_{\min}}$.*

PROOF. Given (w, P), we first compute an estimate $\tilde{r}$ satisfying (2), where $\mu = 2 + \ln\omega$, using $O\big(\frac{\ln\omega}{\sigma^3} \cdot n \ln n\big)$ 1-snapshots. We assume in the sequel that (2) holds. Consider

the following modification of the mixture constituents. We eliminate items i such that $\tilde{r}_i < \frac{2\sigma}{n}$. Each remaining item i is "split" into $n_i = \lfloor n\tilde{r}_i/\sigma \rfloor$ items, and the probability of i is split equally among its copies. The mixture weights are unchanged. From (2), we have that $r_i < \frac{4\sigma}{n}$ if i is eliminated. So the total weight of eliminated items is at most 4σ. Let $n' = \sum_{i:\tilde{r}_i \geq 2\sigma/n} n_i \leq \frac{n}{\sigma}$ be the number of new items. Let $\hat{P} = (\hat{p}^1, \ldots, \hat{p}^k)$ denote the modified mixture constituents, and $\hat{r}$ denote the distribution of the modified 1-snapshots. We prove below that the modified mixture $(w, \hat{P})$ is isotropic and $\zeta/2$-wide.

We use the algorithm for isotropic mixture sources to learn $(w, \hat{P})$ within transportation distance ϵ, using the following procedure to sample m-snapshots from $(w, \hat{P})$. We obtain an m-snapshot from (w, P). We eliminate this snapshot if it includes an eliminated item; otherwise, each item i in the snapshot is replaced by one of its n_i copies, chosen uniformly at random (and independently of previous such choices). From the inferred modified mixture source, we can obtain an estimate of the original mixture source by aggregating, for each inferred mixture constituent, the probabilities of the items that we split, and setting the probability of each eliminated item to 0. This degrades the quality of the solution by the weight of the eliminated items, which is at most an additive $4\sigma \leq \epsilon$ term in the transportation distance.

The probability that an m-snapshot from (w, P) survives is at least $(1 - 4\sigma)^m \geq \frac{1}{2}$ for $m \leq 2k - 1$. Therefore, with probability at least $1 - \frac{1}{3\omega}$, we need at most $6\omega N$ m-snapshots from (w, P) to obtain N m-snapshots from $(w, \hat{P})$. (If we violate this bound, we declare failure.) Thus, we use at most the stated number of 1-, 2-, and $(2k-1)$-snapshots from (w, P) and succeed with probability $1 - O\big(\frac{1}{\omega}\big)$.

We conclude by showing that $(w, \hat{P})$ is isotropic and $\zeta/2$-wide. Let $S = \{i \in [n] : \tilde{r}_i < 2\sigma/n\}$ denote the set of eliminated items. Recall that $\tilde{r}$ satisfies (2). So we have $\frac{31}{32} \leq \frac{\tilde{r}_i}{r_i} \leq \frac{33}{32}$ for every non-eliminated item. We use i_ℓ, where $\ell = 1, \ldots, n_i$, to denote a new item obtained by splitting item i. Define $n_i = 0$ if i is eliminated.

The number n' of new items is at most $\frac{n}{\sigma}$ and at least $\sum_{i \notin S} \frac{2}{3} \cdot \frac{n\tilde{r}_i}{\sigma} \geq \frac{2}{3} \cdot \frac{n}{\sigma} \cdot (1 - 2\sigma) \geq \frac{5n}{8\sigma}$. Let $K = \sum_{i \notin S} r_i \geq 1 - 4\sigma \geq 7/8$. For every new item i_ℓ, we have $\hat{r}_{i_\ell} \geq \frac{r_i}{n\tilde{r}_i/\sigma} \geq \frac{32\sigma}{33n} \geq \frac{1}{2n'}$ and $\hat{r}_{i_\ell} \leq \frac{1}{K} \cdot \frac{3}{2} \cdot \frac{r_i}{n\tilde{r}_i/\sigma} \leq \frac{384\sigma}{217n} \leq \frac{2}{n'}$. Thus, $(w, \hat{P})$ is isotropic.

Now consider the width of $(w, \hat{P})$. For $t = 1, \ldots, k$, define $p'^t \in \mathbb{R}^n$ to be the vector where $p'^t_i = 0$ if $i \in S$, and $p'^t_i = p^t_i$ otherwise. For any distinct $t, t' \in [k]$, we have $\|\hat{p}^t - \hat{p}^{t'}\|_2 \geq \frac{\|\hat{p}^t - \hat{p}^{t'}\|_1}{\sqrt{n'}}$ and

$$\|\hat{p}^t - \hat{p}^{t'}\|_1 = \frac{\|p'^t - p'^{t'}\|_1}{K} \geq \|p^t - p^{t'}\|_1 - \sum_{i \in S} \max\{p^t_i, p^{t'}_i\}$$
$$\geq \zeta - n \cdot \frac{4\sigma}{w_{\min} n} \geq \zeta/2.$$

Let $\hat{A} = \sum_{t=1}^k w_t(\hat{p}^t - \hat{r})(\hat{p}^t - \hat{r})^\dagger$, which is an $n' \times n'$ matrix. We need to prove that the smallest non-zero eigenvalue of $\hat{A}$ is at least $\frac{\zeta^2}{4} \cdot \|\hat{r}\|_\infty$. It will be convenient to define the following matrices. Let $B \in \mathbb{R}^{([n]\setminus S) \times ([n] \setminus S)}$ be the matrix defined by setting $B_{i,j} = A_{i,j}$ for all $i, j \notin S$. Define A' to be the $n \times n$ matrix obtained by padding B with 0s: set $A'_{i,j} = A_{i,j} = B_{i,j}$ if $i, j \notin S$, and equal to 0 otherwise. It is easy to see that the non-zero eigenvalues of A' coincide with the non-zero eigenvalues of B. Define $X \in \mathbb{R}^{n' \times ([n] \setminus S)}$ as follows. Letting $\{i_\ell\}_{i \notin S, \ell = 1, \ldots, n_i}$ index the rows of X, we set $X_{i_\ell, j} = \frac{1}{Kn_i}$ if $j = i$, and 0 otherwise. Notice that $\hat{A} = XBX^\dagger$. To see this, it is convenient to define a padded version $Y \in \mathbb{R}^{n' \times [n]}$ of X by setting $Y_{i_\ell, j} = X_{i_\ell, j}$ if $j \notin S$ and 0 otherwise. Then, we have $\hat{p}^t = Yp^t$ for all $t \in [k]$, and hence, $\hat{A} = YAY^\dagger = XBX^\dagger$.

Note that $\operatorname{rank}(A') \leq \operatorname{rank}(A) = k'$. Consider $A - A'$. Suppose $i \in S$, so $p^t_i \leq \frac{4\sigma}{w_{\min} n}$ for all $t \in [k]$. Then,

$$|(A - A')_{i,j}| = |A_{i,j}| = |M_{i,j} - R_{i,j}| \leq \max\Big\{\sum_{t=1}^k w_t p^t_i p^t_j, r_i r_j\Big\}$$
$$\leq \frac{4\sigma}{w_{\min} n} \cdot r_j \leq \frac{4\sigma}{w_{\min} n} \cdot \|r\|_\infty.$$

Hence, $\|A - A'\|_{\mathrm{op}} \leq \|A - A'\|_F \leq \frac{8\sigma}{w_{\min}} \cdot \|r\|_\infty$. By Lemma 2.1, this implies that

$$\lambda_{k'}(B) = \lambda_{k'}(A') \geq \lambda_{k'}(A) - \|A - A'\|_{\mathrm{op}}$$
$$\geq \Big(\zeta^2 - \frac{8\sigma}{w_{\min}}\Big)\|r\|_\infty \geq \frac{\zeta^2}{2} \cdot \|r\|_\infty.$$

We now argue that $\lambda_{k'}(XBX^\dagger) \geq \lambda_{k'}(B)/(\max_i n_i)$. By the Courant-Fischer theorem (see, e.g., Theorem 4.2.11 in [30]), this is equivalent to showing that there exist vectors $y^1, \ldots, y^{k'} \in \mathbb{R}^{n'}$, such that for every unit vector $v \in \operatorname{Span}(y^1, \ldots, y^{k'})$, we have $v^\dagger(XBX^\dagger)v \geq \frac{\lambda_{k'}(B)}{\max_i n_i}$. We know that there are vectors $u^1, \ldots, u^{k'} \in \mathbb{R}^{[n] \setminus S}$ such that $zBz^\dagger \geq \lambda_{k'}(B)\|z\|_2$ for every $z \in \operatorname{Span}(u^1, \ldots, u^{k'})$. Set $y^t_{i_\ell} = u^t_i$ for every copy i_ℓ of item $i \in [n] \setminus S$, and every $t \in [k']$. Consider any $v \in \operatorname{Span}(y^1, \ldots, y^{k'})$. We have that $z = X^\dagger v \in \operatorname{Span}(u^1, \ldots, u^{k'})$, and since $v_{i_\ell} = z_i$ for every copy i_ℓ of item $i \in [n] \setminus S$ we have that $\|v\|_2^2 \leq (\max_i n_i)\|z\|_2^2$. Therefore, if v is a unit vector, we have $v^\dagger XBX^\dagger v = z^\dagger Bz \geq \lambda_{k'}(B)\|z\|_2^2 \geq \frac{\lambda_{k'}(B)}{\max_i n_i}$.

Putting everything together, we have that $\lambda_{k'}(\hat{A}) \geq \frac{\zeta^2 \|r\|_\infty}{2 \max_i n_i}$. Note that $\|r\|_\infty \geq \frac{32}{33}\|\tilde{r}\|_\infty$ and $\frac{\|\tilde{r}\|_\infty}{\max_i n_i} \geq \frac{\sigma}{n} \geq \frac{217}{384}\|\hat{r}\|_\infty$. So the smallest non-zero eigenvalue of $\hat{A}$ is $\lambda_{k'}(\hat{A}) \geq \frac{\zeta^2}{4}\|\hat{r}\|_\infty$. □

Algorithm overview. Our algorithm for learning an isotropic k-mixture source on $[n]$ takes three parameters: $\zeta \leq 1$ such that (w, P) is ζ-wide, $\omega \in \mathbb{N}$, which controls the success probability of the algorithm, and $\delta \in (0, 1)$, which controls the statistical distance between the constituents of the learnt model and the constituents of the correct model. For convenience, we assume that δ is sufficiently small. The output of the algorithm is a k-mixture source $(\tilde{w}, \tilde{P})$ such that with probability $1 - O\big(\frac{1}{\omega}\big)$, $\|w - \tilde{w}\|_\infty$ and $\|p^t - \tilde{p}^t\|_1$ for all $t \in [k]$ tend to 0 as $\delta \to 0$ (see Theorem 4.1).

The algorithm (see Algorithm 1) consists of three stages. First, we reduce the dimensionality of the problem from n to k' using only 1- and 2-snapshots. By Lemma 3.2, we have an estimate $\tilde{r}$ that is component-wise close to r. Thus, $\tilde{R} = \tilde{r}\tilde{r}^\dagger$ is close in operator norm to R. So we focus on learning the column space of A for which we employ spectral techniques. Leveraging Theorem 2.3, we argue (Lemma 4.2) that by using $O(n \ln^6 n)$ 2-snapshots, one can compute (with high probability) a good enough estimate $\tilde{M}$ of M, and hence obtain a PSD matrix $\tilde{A}$ such that $\|A - \tilde{A}\|_{\mathrm{op}}$ is small.

The remaining task is to learn the projection of P on the affine space $\tilde{r} + \mathrm{Span}(\tilde{A})$, and the mixture weights, which then yields the desired k-mixture source $(\tilde{w}, \tilde{P})$. We divide this into two steps. We choose a random orthonormal basis $\{b_1, \ldots, b_{k'}\}$ of $\mathrm{Span}(\tilde{A})$. For each b_j, we consider the projected k-mixture source $(w, \pi_{b_j}(P))$ on $\mathbb{R}$. In Section 5, we devise a procedure to learn the corresponding k-spike distribution $(w, \mathrm{E}[\pi_{b_j}(P)])$ using $(2k-1)$-snapshots from $(w, \pi_{b_j}(P))$ (which we can obtain using $(2k-1)$-snapshots from (w, P)). Applying this procedure (see Lemma 4.7), we obtain weights $\tilde{w}_1^j, \ldots, \tilde{w}_k^j$ and k (distinct) values $\alpha_1^j, \ldots, \alpha_k^j$ such that each true spike $(w_t, b_j^\dagger p^t)$ maps to a distinct inferred spike $(\tilde{w}^j_{\sigma^j(t)}, \alpha^j_{\sigma^j(t)})$.

Finally, we match up σ_j and $\sigma_{k'}$ for all $j \in [k'-1]$ to obtain k points in $\tilde{r} + \mathrm{Span}(\tilde{A})$ that are close to the projection of P on $\tilde{r} + \mathrm{Span}(\tilde{A})$. For every $j \in [k'-1]$, we generate a random unit "test vector" z_j in $\mathrm{Span}(b_j, b_{k'})$ and learn the projections $\{z_j^\dagger p^t\}_{t\in[k]}$. Since (w, P) is ζ-wide, results about random projections and the guarantees obtained from our k-spike learning procedure imply that $z_j^\dagger(\alpha_{t_1}^j b_j + \alpha_{t_2}^{k'} b_{k'})$ is close to some value in $\{z_j^\dagger p^t\}_{t\in[k]}$ iff there is some t such that $\alpha_{t_1}^j$ and $\alpha_{t_2}^{k'}$ are close respectively to $b_j^\dagger p_t$ and $b_{k'}^\dagger p^t$ (Lemma 4.8). Thus, we can use the learned projections of $\{z_j^\dagger p^t\}_{t\in[k]}$ to match up $\{\alpha_t^j\}_{t\in[k]}$ and $\{\alpha_t^{k'}\}_{t\in[k]}$.

ALGORITHM 1. *Input: an isotropic ζ-wide k-mixture source (w, P) on $[n]$, and parameters $\omega > 1$ and $\delta > 0$.*
Output: a k-mixture source $(\tilde{w}, \tilde{P})$ on $[n]$ that is "close" to (w, P).
Define $T = 3\omega k^4$, $H = \frac{4}{w_{\min}^2 \zeta\sqrt{n}}$ and $L = \frac{\zeta}{64\omega^{1.5}k^4\sqrt{n}}$. We assume that $\delta \le \frac{w_{\min}^3 \zeta^4}{2^{29}\omega^9 k^{16}}$. Let $\kappa = \kappa(2 + \ln\omega)$ be given by Theorem 2.3; we assume $\kappa \ge 1$ for convenience. Define $c = \frac{6400\kappa^2}{w_{\min}^2\delta^2} \cdot \ln\left(\frac{1}{\delta}\right)$. We assume that $w_{\min}^2 \ge \frac{240\kappa \ln^{2.5} n}{\sqrt{n}}$.

A1. **Dimension reduction.**

- A1.1 Use Lemma 3.2 with $\mu = 2 + \ln\omega$ and $\sigma = \frac{\delta}{48}$ to compute an estimate $\tilde{r}$ of r. Set $\tilde{R} = \tilde{r}\tilde{r}^\dagger$.
- A1.2 Independent of all other random variables, choose a Poisson random variable N_2 with expectation $\mathrm{E}[N_2] = cn\ln^6 n$. Choose N_2 independent 2-snapshots and construct a symmetric $n \times n$ matrix $\tilde{M}$ as follows: set $\tilde{M}_{i,i}$ = frequency of the 2-snapshot (i, i) in the sample for all $i \in [n]$, and $\tilde{M}_{i,j} = \tilde{M}_{j,i}$ = half the combined frequency of 2-snapshots (i, j) and (j, i) in the sample, for all $i, j \in [n], i \ne j$.
- A1.3 Compute the spectral decomposition $\tilde{M} - \tilde{R} = \sum_{i=1}^n \lambda_i v_i v_i^\dagger$, where $\lambda_1 \ge \ldots \ge \lambda_n$.
- A1.4 Set $\tilde{A} = \sum_{i:\lambda_i \ge \zeta^2/2n} \lambda_i v_i v_i^\dagger$. Note that $\tilde{A}$ is PSD.

A2. **Learning projections of (w, P) on random vectors in $\mathrm{Span}(\tilde{A})$.**

- A2.1 Pick an orthonormal basis $B = \{b_1, \ldots, b_{k'}\}$ for $\mathrm{Span}(\tilde{A})$ uniformly at random.
- A2.2 Set $(\tilde{w}^j, \alpha^j) \leftarrow \mathsf{Learn}\left(b_j, \delta, \frac{1}{6\omega k}\right)$ for all $j = 1, \ldots, k'$.

A3. **Combining the projections to obtain $(\tilde{w}, \tilde{P})$.**

- A3.1 Pick $\theta \in [0, 2\pi]$ uniformly at random.
- A3.2 For each $j = 1, \ldots, k'-1$, we do the following.
  - Let $z_j = b_j \cos\theta + b_{k'} \sin\theta$.
  - Set $(\hat{w}^j, \hat{\alpha}^j) \leftarrow \mathsf{Learn}\left(z_j, \delta, \frac{1}{6\omega k}\right)$.
  - For each $t_1, t_2 \in [k]$, if there exists $t \in [k]$ such that $\left|(\alpha_{t_1}^j b_j + \alpha_{t_2}^{k'} b_{k'})^\dagger z_j - \hat{\alpha}_t^j\right| \le (\sqrt{2}+1)L/(2+5T)$ then set $\varrho^j(t_2) = t_1$.
- A3.3 Define $\varrho^{k'}(t) = t$ for all $t \in [k]$.
- A3.4 For every $t \in [k]$: set $\tilde{w}_t = \left(\sum_{j=1}^{k'} \tilde{w}^j_{\varrho^j(t)}\right)/k'$, $\hat{p}^t = \tilde{r} + \sum_{j=1}^{k'}\left(\alpha^j_{\varrho^j(t)} - b_j^\dagger \tilde{r}\right) b_j$, and $\tilde{p}^t = \arg\min_{x\in\Delta^{n-1}} \|x - \hat{p}^t\|_1$, which can be computed by solving an LP. Return $\left(\tilde{w}, \tilde{P} = (\tilde{p}^1, \ldots, \tilde{p}^k)\right)$.

ALGORITHM $\mathsf{Learn}(v, \varsigma, \varepsilon)$
Input: a unit vector $v \in \mathrm{Span}(\tilde{A})$, and parameters $\varsigma > 0$, $\varepsilon > 0$. We assume that (a) $|v^\dagger(p-q)| \ge L$ for all distinct $p, q \in P$; and (b) $1024k\varsigma < \frac{w_{\min} L}{16H}$.
Output: a k-spike distribution $\left(\bar{w}, (\gamma_1, \ldots, \gamma_k)\right)$ close to $(w, \mathrm{E}[\pi_v(P)])$.

- L1. Solve the following convex program:

$$\min \quad \|x\|_\infty \quad \text{s.t.} \quad v^\dagger x \ge 1 - \frac{4\delta}{\zeta^2}, \quad \|x\|_2^2 \le 1 \qquad (\mathrm{Q}_v)$$

  to obtain x^*; set $a = \frac{x^*}{\|x^*\|_2}$. We prove in Lemma 4.4 that $\|a\|_\infty \le H$ and $|a^\dagger(p-q)| \ge \frac{L}{2}$ for all $p, q \in P,\ p \ne q$.
- L2. Let $s = \varsigma^{4k}$. Apply the procedure in Section 5 leading to Theorem 5.1 for $\left(w, \pi_{a/2H}(P)\right)$ to infer a k-spike distribution $(\bar{w}, \beta)$ that, with probability at least $1 - \varepsilon$, is within transportation distance $O\left(s^{\Omega(1/k)}\right)$ from $\left(w, \mathrm{E}[\pi_{a/2H}(P)]\right)$. This uses a sample of $(2k-1)$-snapshots of size $3k2^{4k}s^{-4k}\ln(4k/\varepsilon)$.
- L3. For every $t \in [k]$, set $\gamma_t = (2H\beta_t)(a^\dagger v)$. Return $(\bar{w}, \gamma)$.

Remark 1. We cannot compute the spectral decomposition in step A1.3 exactly, or solve (Q_v) exactly in step L1, since the output may be irrational. However, one can obtain a decomposition such that $\|\tilde{M} - \tilde{R} - \sum_{i=1}^n \lambda_i v_i v_i^\dagger\|_{\mathrm{op}} = O\left(\frac{\delta}{n}\right)$ and compute a 2-approximate solution to (Q_v) in polytime, and this suffices: slightly modifying the constants H and c makes the entire analysis go through. We have chosen the presentation above to keep exposition simple.

---

4. ANALYSIS

THEOREM 4.1. *Algorithm 1 uses $O\left(\frac{\ln\omega}{\delta^3} \cdot n\ln n\right)$ 1-snapshots, $O\left(\frac{\ln^2\omega\ln(1/\delta)}{\delta^2 w_{\min}^2} \cdot n\ln^6 n\right)$ 2-snapshots, and $O\left(\frac{k2^{4k}}{\delta^{16}k^2} \cdot \ln(24\omega k^2)\right)$ $(2k-1)$-snapshots, and computes a k-mixture source $(\tilde{w}, \tilde{P})$ on $[n]$ such that with probability $1 - O\left(\frac{1}{\omega}\right)$, there is a permutation $\sigma : [k] \mapsto [k]$ such that for all $t = 1, \ldots, k$,*

$$|w_t - \tilde{w}_{\sigma(t)}| = O\left(\frac{\delta\omega^{1.5}k^5}{w_{\min}^2\zeta^2}\right) \quad \text{and} \quad \|p^t - \tilde{p}^{\sigma(t)}\|_1 = O\left(\frac{\sqrt{k\delta}}{w_{\min}^{1.5}\zeta}\right).$$

Hence, $\mathrm{Tran}(w, P; \tilde{w}, \tilde{P}) = O\left(\frac{\sqrt{k\delta}}{w_{\min}^{1.5}\zeta}\right)$. *The running time is polynomial in the sample size.*

The roadmap of the proof is as follows. By Lemma 3.2, with probability at least $1 - \frac{1}{\omega n^2}$, $\left(1 - \frac{\delta}{48}\right)r_i \le \tilde{r}_i \le \left(1 + \frac{\delta}{48}\right)r_i$ for all $i \in [n]$. We assume that this holds in the sequel. In Lemma 4.2, we prove that the matrix $\tilde{A}$ computed after step A1 is a good estimate of A. In Lemma 4.3, we derive some properties of the column space of A. Lemma 4.4 then uses these properties to show that algorithm Learn returns a good approximation to $(w, \mathrm{E}[\pi_v(P)])$. Claim 4.5 and Lemma 4.6 prove that the projections of the mixture constituents on the b_js and the z_js are well-separated. Combining this with Lemma 4.4, we prove in Lemma 4.7 that with suitably large probability, every true spike $(w_t, b_j^\dagger p^t)$ maps to a distinct nearby inferred spike on every b_j, $j \in [k']$, and similarly

every true spike $(w_t, z_j^\dagger p^t)$ maps to a distinct nearby inferred spike on every $z_j,\ j \in [k'-1]$. Lemma 4.8 shows that one can then match up the spikes on the different b_js. This yields k points in $\mathrm{Span}(\tilde{A})$ that are close to the projection of P on $\mathrm{Span}(\tilde{A})$. Finally, we argue that this can be mapped to a k-mixture source $(\tilde{w}, \tilde{P})$ that is close to (w, P).

LEMMA 4.2. *With probability at least* $1 - \frac{1}{n\omega}$, *the matrix* $\tilde{A}$ *computed after step A1 satisfies* $\mathrm{rank}(\tilde{A}) = k' = \mathrm{rank}(A)$ *and* $\|A - \tilde{A}\|_{\mathrm{op}} \le \frac{\delta}{n}$.

PROOF. Recall that $k' = \mathrm{rank}(A)$. Let $B = \tilde{M} - \tilde{R} = \sum_{i=1}^n \lambda_i v_i v_i^\dagger$, where $\lambda_1 \ge \ldots \ge \lambda_n$. We prove below that with probability at least $1 - \frac{1}{n\omega}$, we have $\|M - \tilde{M}\|_{\mathrm{op}} \le \frac{\delta}{4n}$ and $\|R - \tilde{R}\|_{\mathrm{op}} \le \frac{\delta}{4n}$. This implies that $\|A - B\|_{\mathrm{op}} \le \|M - \tilde{M}\|_{\mathrm{op}} + \|R - \tilde{R}\|_{\mathrm{op}} \le \frac{\delta}{2n}$. Hence, by Lemma 2.1, it follows that by the ζ-wide assumption, $\lambda_{k'} \ge \frac{\zeta^2}{n} - \frac{\delta}{2n} \ge \frac{3\zeta^2}{4n}$, and $|\lambda_i| \le \frac{\delta}{2n} \le \frac{\zeta^2}{4n}$ for all $i > k'$. Thus, we include exactly k' eigenvectors when defining $\tilde{A}$, so $\mathrm{rank}(\tilde{A}) = k'$. Since $\tilde{A}$ is the closest rank-k' approximation in operator norm to B, we have $\|A - \tilde{A}\|_{\mathrm{op}} \le \|A - B\|_{\mathrm{op}} + \|B - \tilde{A}\|_{\mathrm{op}} \le 2\|A - B\|_{\mathrm{op}} \le \frac{\delta}{n}$.

It is easy to see that $|\tilde{R}_{i,j} - R_{i,j}| \le 3\sigma r_{i,j}$, where $\sigma = \delta/48$, and so $\|R - \tilde{R}\|_{\mathrm{op}} \le \|R - \tilde{R}\|_F \le \frac{\delta}{4n}$. Bounding $\|M - \tilde{M}\|_{\mathrm{op}}$ is more challenging. We carefully define a matrix whose entries are independent random variables with bounded variance, and then apply Theorem 2.3.

Note that $M_{i,j} \le \min\{\frac{2}{n}, \frac{4}{w_{\min} n^2}\}$ due to isotropy. Let $K = \frac{4\ln(1/\delta)}{\delta}$ and $K' = \frac{5\ln(1/\delta)}{\delta}$. Let $D = N_2 \cdot (\tilde{M} - M)$. Let $X_{i,i}^\ell = 1$ if the ℓ-th snapshot is (i,i), for $i \in [n]$, and for $i, j \in [n], i \ne j$, let $X_{i,j}^\ell = X_{j,i}^\ell = \frac{1}{2}$ if the ℓ-th 2-snapshot is (i,j) or (j,i), and 0 otherwise. Let $Y_{i,j}^\ell = X_{i,j}^\ell - M_{i,j} = X_{i,j}^\ell - \mathrm{E}[X_{i,j}^\ell]$; so $D_{i,j} = \sum_{\ell=1}^{N_2} Y_{i,j}^\ell$ for all $i, j \in [n]$. We have $\sigma^2(n_2) := \mathrm{Var}[D_{i,j} | N_2 = n_2] = n_2 \mathrm{Var}[X_{i,j}^1] \le n_2 \mathrm{E}[(X_{i,j}^1)^2] \le n_2 M_{i,j}$. For $n_2 \le 2cn\ln^6 n$, we have $\sigma^2(n_2) \le \frac{8c\ln^6 n}{w_{\min}^2 n} \le \frac{\ln n \ln(1/\delta)}{\delta^2}$ (since $w_{\min}^4 \ge \frac{57600\kappa^2 \ln^5 n}{n}$). So by Bernstein's inequality (Lemma A.2),

$$\Pr[|D_{i,j}| > K\ln n | N_2 = n_2] \le 2\exp\Big(-\frac{K^2 \ln^2 n}{2\big(\sigma^2(n_2) + K\ln n/3\big)}\Big)$$
$$\le 2\max\Big\{\exp\big(-\tfrac{K^2\ln^2 n}{4\sigma^2(n_2)}\big), \exp\big(-\tfrac{3K\ln n}{4}\big)\Big\} \le \frac{2\delta}{n^3}.$$

Since $\Pr[N_2 > 2c\ln^6 n] \le n^{-3}$, we can say that with probability at least $1 - 2n^{-2}$, we have $|D_{i,j}| \le K\ln n$ for every $i, j \in [n]$ and $N_2 \le 2c\ln^6 n$.

Define a matrix D' by putting, for every $i, j \in [n]$, $D'_{i,j} = \mathrm{sign}(D_{i,j}) \cdot \min\{|D_{i,j}|, K\ln n\}$. Put $D'' = D' - \mathrm{E}[D']$. Clearly, $\mathrm{E}[D''_{i,j}] = 0$ for every $i, j \in [n]$. We prove below that $\big|\mathrm{E}[D'_{i,j}]\big| \le \frac{3\delta c\ln^6 n}{n^2} \le \frac{\ln n \ln(1/\delta)}{\delta}$; therefore, $|D''_{i,j}| \le K'\ln n$. The entries of D are independent random variables as N_2 is a Poisson random variable; hence, the entries of D'' are also independent random variables. Also $\mathrm{Var}[D''_{i,j}] \le \mathrm{Var}[D_{i,j}]$ since censoring a random variable to an interval can only reduce the variance. Note that $D_{i,j} = \sum_{\ell=1}^{N_2} Y_{i,j}^\ell$ follows the compound Poisson distribution. So we have

$$\mathrm{Var}[D_{i,j}] = \mathrm{E}[N_2] \cdot \mathrm{E}[(Y_{i,j}^1)^2] = \mathrm{E}[N_2] \cdot \mathrm{Var}[X_{i,j}^1]$$
$$\le \mathrm{E}[N_2] M_{i,j} \le \frac{4c\ln^6 n}{w_{\min}^2 \cdot n} \le \frac{\hat{c}^2 K'^2 \ln^6 n}{n}$$

where $\hat{c} = \max\{\frac{2\sqrt{c}}{w_{\min} K'}, \kappa^2\}$. Thus, by Theorem 2.3, the constant $\kappa = \kappa(2 + \ln\omega) > 0$ is such that with probability at least $1 - \frac{1}{n^2\omega}$

$$\|D''\|_{\mathrm{op}} \le 2 \cdot \frac{\hat{c}K'\ln^3 n}{\sqrt{n}} \cdot \sqrt{n} + \kappa\sqrt{K'\ln n \cdot \frac{\hat{c}K'\ln^3 n}{\sqrt{n}}} \cdot \sqrt[4]{n} \cdot \ln n$$
$$\le \big(2K'\hat{c} + \kappa K'\sqrt{\hat{c}}\big)\ln^3 n. \qquad (3)$$

We have $\Pr\big[N_2 \ge \frac{1}{2}\mathrm{E}[N_2]\big] \ge 1 - n^{-2}$, Thus, with probability at least $1 - \frac{1}{n\omega}$, we have that $N_2 \ge \frac{1}{2}\mathrm{E}[N_2]$, $D' = D$, and $\|D''\|_{\mathrm{op}}$ is bounded by (3). We show below that $2\|\mathrm{E}[D']\|_{\mathrm{op}}/\mathrm{E}[N_2] \le 6\delta n^{-2} \le \delta/20n$. One can verify that $4K'\hat{c}/c \le \delta/10$ and $2\kappa K'\sqrt{\hat{c}}/c \le \delta/10$. Therefore, with probability at least $1 - \frac{1}{n\omega}$, we have that $\|M - \tilde{M}\|_{\mathrm{op}} = \frac{1}{N_2} \cdot \|D\|_{\mathrm{op}} \le \frac{2}{\mathrm{E}[N_2]} \cdot (\|D''\|_{\mathrm{op}} + \|\mathrm{E}[D']\|_{\mathrm{op}}) \le \frac{\delta}{4n}$.

Finally, we bound $\|\mathrm{E}[D']\|_{\mathrm{op}}$. We have $\|\mathrm{E}[D']\|_{\mathrm{op}} \le \|\mathrm{E}[D']\|_F \le n \cdot \max_{i,j}\big|\mathrm{E}[D'_{i,j}]\big|$. Let $\mu = cn\ln^6 n = \mathrm{E}[N_2]$. Fix any i, j. We have $\big|\mathrm{E}[D'_{i,j}]\big| = \big|\mathrm{E}[D'_{i,j} - D_{i,j}]\big| \le \mathrm{E}\big[|D'_{i,j} - D_{i,j}|\big]$. For any $n_2 \le 2\ln(1/\delta)\mu$, we have $\mathrm{Var}[D_{i,j} | N_2 = n_2] \le n_2 M_{i,j} \le \frac{8c\ln(1/\delta)\ln^6 n}{w_{\min}^2 n}$. So by Bernstein's inequality, we have that $\Pr[|D_{i,j}| > K\ln n | N_2 \le 2\ln(1/\delta)\mu] < 2\delta n^{-3}$. Also, $|D'_{i,j} - D_{i,j}| \le N_2$ always. Therefore,

$$\mathrm{E}\big[|D'_{i,j} - D_{i,j}| \big| N_2 = n_2\big] \le \begin{cases} 2\delta n^{-3} n_2 & \text{if } n_2 \le 2\ln\big(\frac{1}{\delta}\big)\mu; \\ n_2 & \text{otherwise} \end{cases}$$

and $\mathrm{E}\big[|D'_{i,j} - D_{i,j}|\big] \le \mu - \Pr[N_2 \le 2\ln(1/\delta)\mu]\,\mathrm{E}[N_2 | N_2 \le 2\ln(1/\delta)\mu](1 - 2\delta n^{-3})$. Since N_2 is Poisson distributed, we have

$$\Pr[N_2 \le 2\ln(1/\delta)\mu]\,\mathrm{E}[N_2 | N_2 \le 2\ln(1/\delta)\mu]$$
$$= \sum_{\ell=0}^{\lfloor 2\ln(1/\delta)\mu \rfloor} \ell \cdot \frac{\mu^\ell e^{-\mu}}{\ell!}$$
$$= \mu \sum_{\ell=0}^{\lfloor 2\ln(1/\delta)\mu \rfloor - 1} \frac{\mu^\ell e^{-\mu}}{\ell!}$$
$$\ge \mu \Pr[N_2 \le \ln(1/\delta)\mu] \ge \mu(1 - \delta n^{-3}).$$

Thus, $\mathrm{E}\big[|D'_{i,j} - D_{i,j}|\big] \le \mu - \mu(1 - \delta n^{-3})(1 - 2\delta n^{-3}) \le 3\delta n^{-3}\mu$, and $2\|\mathrm{E}[D']\|_{\mathrm{op}}/\mathrm{E}[N_2] \le 6\delta n^{-2}$. ☐

We assume in the sequel that the high-probability event stated in Lemma 4.2 happens. Thus, Lemma 2.2 implies that $\|\Pi_A - \Pi_{\tilde{A}}\|_{\mathrm{op}} \le \frac{2\sqrt{\delta}}{\zeta}$.

LEMMA 4.3. *For every unit vector* $b \in \mathrm{Span}(A)$, $\|b\|_\infty \le \frac{2}{w_{\min}^2 \zeta\sqrt{n}}$.

PROOF. Recall that $A = \sum_{t=1}^k w_t (p^t - r)(p^t - r)^\dagger$, and the smallest non-zero eigenvalue of A is at least ζ^2/n. Note that $\mathrm{Span}(A) = \mathrm{Span}\{p^1 - r, \ldots, p^k - r\}$. Let $Z = \mathrm{conv}(P)$. If $r + b \in Z$, then $\|r + b\|_\infty \le \frac{2}{w_{\min} n}$, $r + b \ge 0$, and $\|r\|_\infty \le \frac{2}{n}$ imply that $\|b\|_\infty \le \frac{2}{w_{\min} n}$. Otherwise, let the line segment $[r, r + b]$ intersect the boundary of Z at some point b'. We show that $\|r - b'\|_2^2 \ge \frac{\zeta^2 w_{\min}^2}{n}$. The lemma then follows since $b = (b' - r)/\|b' - r\|_2$ and so $\|b\|_\infty = \frac{\|b' - r\|_\infty}{\|b' - r\|_2} \le \frac{2}{w_{\min}^2 \zeta\sqrt{n}}$.

Let S be a facet of Z such that $b' \in S$, $r \notin S$ (note that r is in the strict interior of P). Since $Z \subseteq \mathrm{Span}(A)$, one can

find a unit vector $v \in \mathrm{Span}(A)$ such that S is exactly the set of points that minimize $v^\dagger x$ over $x \in Z$. Let $d_L = v^\dagger r - \min_{x\in Z} v^\dagger x = v^\dagger(r - b')$. We lower bound $\|r - b'\|_2$ by d_L. Note that $d_L > 0$. Clearly, $v^\dagger(p^t - r) \geq -d_L$ for all $t \in [k]$. Projecting P onto v, we have that (a) $\sum_{t=1}^k w_t v^\dagger(p^t - r) = 0$; and (b) $v^T A v = \sum_{t=1}^k w_t \left(v^\dagger(p^t - r)\right)^2 \geq \frac{\zeta^2}{n}$ since $v \in \mathrm{Span}(A)$ and (w, P) is ζ-wide. Let $W_L = \sum_{t:v^\dagger(p^t-r)\leq 0} w_t$, let $W_R = 1 - W_L \geq w_{\min}$, and let $d_R = \max_t\{v^\dagger(p^t - r)\}$. Then, $0 = \sum_{t=1}^k w_t v^\dagger(p^t - r) \geq W_L(-d_L) + w_{\min} d_R$, so $d_R \leq d_L \cdot \frac{W_L}{w_{\min}}$, Also $\frac{\zeta^2}{n} \leq \sum_{t=1}^k w_t \left(v^\dagger(p^t - r)\right)^2 \leq W_L \cdot d_L^2 + W_R \cdot d_R^2 \leq W_L \cdot d_L^2 + W_R \cdot d_L^2 \cdot \frac{W_L^2}{w_{\min}^2} \leq \frac{d_L^2}{w_{\min}^2}$. So, $d_L^2 \geq \frac{\zeta^2 w_{\min}^2}{n}$. $\square$

LEMMA 4.4. *If the assumptions stated in Algorithm* Learn *are satisfied, then: (i) the vector a computed in* Learn *satisfies $\|a\|_\infty \leq H$, and $|a^\dagger(p-q)| \geq L/2$ for every two mixture constituents $p, q \in P$;*
(ii) with probability at least $1-\varepsilon$, the output $(\bar{w}, \gamma)$ of Learn *satisfies the following: there is a permutation $\sigma : [k] \mapsto [k]$ such that for all $t = 1, \ldots, k$,*

$$|w_t - \bar{w}_{\sigma(t)}| = O\Big(\frac{\varsigma\omega^{1.5}k^5}{w_{\min}^2\zeta^2}\Big), \qquad \textit{and}$$

$$|v^\dagger p^t - \gamma_{\sigma(t)}| \leq \frac{2048kH\varsigma}{w_{\min}} + \frac{8\sqrt{2\delta}}{w_{\min}\zeta\sqrt{n}} \leq \frac{2048kH\varsigma}{w_{\min}} + \frac{L}{8T}.$$

PROOF. We have $v^\dagger \Pi_A(v) = 1 - \|v - \Pi_A(v)\|_2^2 = 1 - \|(\Pi_{\tilde{A}} - \Pi_A)v\|_2^2 \geq 1 - \frac{4\delta}{\zeta^2}$. Thus, $\Pi_A(v)$ is feasible to (Q_v), and since $\|\Pi_A(v)\|_2 \leq 1$, by Lemma 4.3, the optimal solution x^* to (Q_v) satisfies $\|x^*\|_\infty \leq 2\|\Pi_A(v)\|_\infty \leq H/2$. Also $\|x^*\|_2^2 \geq v^\dagger x^* \geq 1 - \frac{4\delta}{\zeta^2} \geq \frac{1}{4}$, so $\|a\|_\infty \leq H$. Note that $\|v - a\|_2^2 = 2(1 - v^\dagger a) \leq 2(1 - v^\dagger x^*) \leq \frac{8\delta}{\zeta^2}$. It follows that for any two mixture constituents p, q, we have

$$\begin{aligned}
|a^\dagger(p-q)| &\geq |v^\dagger(p-q)| - |(v-a)^\dagger(p-q)| \\
&\geq |v^\dagger(p-q)| - \frac{2\sqrt{2\delta}}{\zeta}\|p-q\|_2 \\
&\geq |v^\dagger(p-q)| - \frac{8\sqrt{2\delta}}{w_{\min}\zeta\sqrt{n}} \geq |v^\dagger(p-q)| - \frac{L}{2} \geq \frac{L}{2}.
\end{aligned}$$

This proves part (i). For part (ii), we note that any two spikes in the k-spike mixture $\big(w, \mathrm{E}[\pi_{a/2H}(P)]\big)$ are separated by a distance of at least $L/4H$. Since $s < L/4H$, Theorem 5.1 guarantees that with a sample of $(2k-1)$-snapshots of size $3k2^{4k}s^{-4k}\log(4k/\varepsilon)$, with probability at least $1-\varepsilon$, the learned k-spike distribution $(\bar{w}, \beta)$ satisfies $\mathrm{Tran}\big(w, \mathrm{E}[\pi_{a/2H}(P)]; \bar{w}, \beta\big) \leq 1024ks^{1/(4k)} = 1024k\varsigma < \frac{Lw_{\min}}{8H}$. Notice that this implies that there is a permutation $\sigma : [k] \mapsto [k]$ such that $\forall t = 1, \ldots, k$:

$$|(a/2H)^\dagger p^t - \beta_{\sigma(t)}| \leq \frac{1024k\varsigma}{w_{\min}} < \frac{L}{8H}, \tag{4}$$

$$|w_t - \bar{w}_{\sigma(t)}| = O\Big(\frac{k\varsigma}{L/8H}\Big) = O\Big(\frac{\varsigma\omega^{1.5}k^5}{w_{\min}^2\zeta^2}\Big).$$

Fix some $t \in [k]$. Let $t' = \sigma(t)$. From (4), we know that $|a^\dagger p^t - 2H \cdot \beta_{t'}| = \frac{2048kH\varsigma}{w_{\min}}$. We bound $|v^\dagger p^t - a^\dagger p^t|$ and $|2H\beta_{t'} - \gamma_{t'}|$, which together with the above will complete the proof of the lemma. We have $|(v-a)^\dagger p^t| \leq \|v - a\|_2\|p^t\|_2 \leq \frac{4\sqrt{2\delta}}{w_{\min}\zeta\sqrt{n}}$. Since $\gamma_{t'} = (2H\beta_{t'})a^\dagger v$ and $|\beta_{t'}| \leq \frac{1}{2}$, we have $|2H\beta_{t'} - \gamma_{t'}| \leq \frac{H\cdot 4\delta}{\zeta^2} \leq \frac{16\delta}{w_{\min}^2\zeta^3\sqrt{n}}$. It follows that $|v^\dagger p^t - \gamma_{t'}| \leq \frac{2048kH\varsigma}{w_{\min}} + \frac{8\sqrt{2\delta}}{w_{\min}\zeta\sqrt{n}} \leq \frac{2048kH\varsigma}{w_{\min}} + \frac{L}{8T}$. $\square$

CLAIM 4.5. *Let Z be a random unit vector in* $\mathrm{Span}(\tilde{A})$ *and $v \in \mathrm{Span}(\tilde{A})$. $\Pr\big[|Z^\dagger v| < \frac{\|v\|_2}{32\omega^{1.5}k^4}\big] < \frac{1}{3\omega k' k^2}$.*

PROOF. One way of choosing the random unit vector Z is as follows. Fix an orthonormal basis $\{u_1, \ldots, u_{k'}\}$ for $\mathrm{Span}(\tilde{A})$. We choose independent $N(0,1)$ random variables X_i for $i \in [k']$. Define $C = \sum_{i=1}^{k'} X_i u_i$ and set $Z = C/\|C\|_2$. Set $a_1 = \frac{\pi}{32\omega^2 k'^2 k^4}$ and $a_2 = 2 + \frac{4\ln(12\omega k' k^2)}{k'} \leq 96\omega k$.

Note that $C^\dagger v/\|v\|_2$ is distributed as $N(0,1)$. Therefore, $\Pr\big[|C^\dagger v| \leq \|v\|_2\sqrt{a_1}\big] \leq \sqrt{\frac{2a_1}{\pi}} \leq \frac{1}{4\omega k' k^2}$. Also, $\|C\|_2^2 = \sum_{i=1}^{k'} X_i^2$ follows the $\chi^2_{k'}$ distribution. So

$$\begin{aligned}
\Pr[\|C\|_2^2 > a_2 k'] &< \Big(a_2 e^{1-a_2}\Big)^{k'/2} \\
&< \exp\Big((1-a_2/2)k'/2\Big) < \frac{1}{12\omega k' k^2}.
\end{aligned}$$

Observe that $\sqrt{\frac{a_1}{a_2 k'}} \geq \frac{1}{32\omega^{1.5}k^4}$. So if the "bad" event stated in the lemma happens, then $|C^\dagger v| \leq \|v\|_2\sqrt{a_1}$ or $\|C\|_2^2 \geq a_2 k'$ happens; the probability of this is at most $\frac{1}{3\omega k' k^2}$. $\square$

LEMMA 4.6. *With probability at least $1 - \frac{1}{3\omega}$, for every pair $p, q \in P$, we have (i) $|b_j^\dagger(p-q)| \geq L$ for every $j \in [k']$ and (ii) $|z_j^\dagger(p-q)| \geq L$ for every $j \in [k'-1]$.*

PROOF. Define $\tilde{p} = \Pi_{\tilde{A}}(p)$ for a mixture constituent p. Clearly, for any $v \in \mathrm{Span}(\tilde{A})$, $v^\dagger\tilde{p} = v^\dagger p$. Recall that $\|\Pi_A - \Pi_{\tilde{A}}| \leq \frac{2\sqrt{\delta}}{\zeta}$. So for every $p, q \in P$, $\|\tilde{p} - \tilde{q}\|_2^2 \geq \|p - q\|_2^2 - \|(\Pi_A - \Pi_{\tilde{A}})(p-q)\|_2^2 \geq \|p-q\|^2/4$; hence, $\|\tilde{p} - \tilde{q}\|_2 \geq \frac{\zeta}{2\sqrt{n}}$. Notice that the z_j vectors are also random unit vectors in $\mathrm{Span}(\tilde{A})$. Applying Claim 4.5 to each event involving one of the $\{b_j\}_{j\in[k']}, \{z_j\}_{j\in[k'-1]}$ random unit vectors, and one of the $\binom{k}{2}$ vectors $\|\tilde{p} - \tilde{q}\|$ for $\tilde{p}, \tilde{q} \in \Pi_{\tilde{A}}(P)$, and taking the union bound over the at most $k'k^2$ such events completes the proof. $\square$

LEMMA 4.7. *With probability at least $1 - \frac{2}{3\omega}$, the k-spike distributions obtained in steps A2 and A3 satisfy:*
(i) For every $j \in [k']$, there is a permutation $\sigma^j : [k] \mapsto [k]$ such that for all $t \in [k]$,

$$|w_t - \tilde{w}^j_{\sigma^j(t)}| = O\Big(\frac{\delta\omega^{1.5}k^5}{w_{\min}^2\zeta^2}\Big),$$

$$|b_j^\dagger p^t - \alpha^j_{\sigma^j(t)}| = O\Big(\frac{\sqrt{\delta}}{w_{\min}^{1.5}\zeta\sqrt{n}}\Big) \quad \textit{and is at most } \frac{L}{2+5T}.$$

Hence, $|\alpha^j_{t_1} - \alpha^j_{t_2}| \geq L - \frac{2L}{2+5T} = \frac{L}{1+0.4/T}$ for all distinct $t_1, t_2 \in [k]$.
(ii) For every $j \in [k'-1]$, for every $t \in [k]$, there is a distinct t' such that

$$|w_t - \hat{w}^j_{t'}| = O\Big(\frac{\delta\omega^{1.5}k^5}{w_{\min}^2\zeta^2}\Big),$$

$$|z_j^\dagger p^t - \hat{\alpha}^j_{t'}| = O\Big(\frac{\sqrt{\delta}}{w_{\min}^{1.5}\zeta\sqrt{n}}\Big) \quad \textit{and is at most } \frac{L}{2+5T}.$$

PROOF. Assume that the event stated in Lemma 4.6 happens. Then the inputs to Learn in steps A2 and A3 are

"valid", i.e., satisfy the assumptions stated in Algorithm Learn. Plug in $\varsigma = \delta$ and $\varepsilon = \frac{1}{6\omega k}$ in Lemma 4.4. Taking the union bound over all the b_js and the z_js, we obtain that the probability that Learn fails on some input, when all the b_js and z_js are valid is at most $\frac{1}{3\omega}$. The lemma follows from Lemma 4.4 by noting that $\frac{2048kH\delta}{w_{\min}} = O\Big(\frac{\sqrt{\delta}}{w_{min}^{1.5}\zeta\sqrt{n}}\Big)$ and is at most $\frac{L}{24T}$, and $L/24T + L/8T \le L/(2+5T)$. $\square$

LEMMA 4.8. *With probability at least $1-\frac{1}{\omega}$, for every $j = 1,\ldots,k'-1$ ϱ^j is a well-defined function and $\varrho^j(\sigma^{k'}(t)) = \sigma^j(t)$ for every $t \in [k]$.*

PROOF. Assume that the events in Lemmas 4.6 and 4.7 occur. Fix $j \in [k'-1]$. We call a point $\alpha^j_{t_1} b_j + \alpha^{k'}_{t_2} b_{k'}$ a grid-j point. Call this grid point "genuine" if there exists $t \in [k]$ such that $\sigma^j(t) = t_1$ and $\sigma^{k'}(t) = t_2$, and "fake" otherwise. The distance between any two grid-j points is at least $L/(1+0.4/T)$ (by Lemma 4.7). So the probability there is a pair of genuine and fake grid-j points whose projections on z_j are less than $L/(T+0.4)$ away is at most $k^3 \cdot \frac{2}{\pi}\arcsin\big(\frac{1}{T}\big) \le k^3 \cdot \frac{2}{\pi} \cdot \frac{6}{5T} \le \frac{1}{3\omega k}$. Therefore, with probability at least $1-\omega$, the events in Lemma 4.6 and Lemma 4.7 happen, and for all $j \in [k'-1]$, every pair of genuine and fake grid-j points project to points on z_j that are at least $L/(T+0.4)$ apart. We condition on this in the sequel.

Now fix $j \in [k'-1]$ and consider any pair $t_1, t_2 \in [k]^2$. Let g be the grid-j point $b_j\alpha^j_{t_1} + b_{k'}\alpha^{k'}_{t_2}$ We show that $\varrho^j(t_2) = t_1$ iff g is a genuine grid-j point. If g is genuine, let t be such that $\sigma^j(t) = t_1$, $\sigma^{k'}(t) = t_2$. Let p' be the projection of p^t on $\mathrm{Span}(b_j, b_{k'})$. By Lemma 4.7, we have that $\|p'-g\|_2 \le \frac{\sqrt{2}L}{2+5T}$. Also, there exists $t' \in [k]$ such that $|\hat{\alpha}^j_{t'} - z_j^\dagger p^t| \le \frac{L}{2+5T}$. Since $z_j^\dagger p' = z_j^\dagger p^t$, this implies that $|z_j^\dagger g - \hat{\alpha}^j_{t'}| \le |\hat{\alpha}^j_{t'} - z_j^\dagger p^t| + |z_j^\dagger(p'-g)| \le \frac{(\sqrt{2}+1)L}{2+5T}$ and so $\varrho^j(t_2) = t_1$.

Now suppose g is fake but $|z_j^\dagger g - \hat{\alpha}^j_{t'}| \le (\sqrt{2}+1)L/(2+5T)$ for some $t' \in [k]$. Let $t \in [k]$ be such that $|\hat{\alpha}^j_{t'} - z_j^\dagger p^t| \le \frac{L}{2+5T}$. Let g' be the genuine grid point $b_j\alpha^j_{\sigma^j(t)} + b_{k'}\alpha^{k'}_{\sigma^{k'}(t)}$. So $|z_j^\dagger g' - \hat{\alpha}^j_{t'}| \le (\sqrt{2}+1)L/(2+5T)$, and hence $|z_j^\dagger(g-g')| \le \frac{2(\sqrt{2}+1)L}{2+5T} < \frac{L}{0.4+T}$ which is a contradiction. $\square$

PROOF OF THEOREM 4.1. We condition on the fact that all the "good" events stated in Lemmas 3.2, 4.2, 4.6, 4.7, and 4.8 happen. The probability of success is thus $1 - O\big(\frac{1}{\omega}\big)$. The sample-size bounds follow from the description of the algorithm. For notational simplicity, let $\sigma^{k'}$ be the identity permutation, i.e., $\sigma^{k'}(t) = t$ for all $t \in [k]$. So by Lemma 4.8, we have $\varrho^j(t) = \sigma^j(t)$ for every $j \in [k'-1]$ and $t \in k$.

For $t = 1,2,\ldots,k$, define $\bar{p}^t = \tilde{r} + \sum_{j=1}^{k'} b_j^\dagger(p^t - \tilde{r})b_j = \tilde{r} + \Pi_{\tilde{A}}(p^t - \tilde{r})$. Fix $t \in [k]$. Then

$$\begin{aligned}\|p^t - \tilde{p}^t\|_1 &\le \|p^t - \hat{p}^t\|_1 + \|\hat{p}^t - \tilde{p}^t\|_1 \\ &\le 2\|p^t - \hat{p}^t\|_1 \le 2\left(\|p^t - \bar{p}^t\|_1 + \|\bar{p}^t - \hat{p}^t\|_1\right).\end{aligned}$$

We have

$$\begin{aligned}\|p^t - \bar{p}^t\|_2 &= \Big\|r - \tilde{r} + (p^t - r) - \sum_{j=1}^{k'} b_j^\dagger(p^t - \tilde{r})b_j\Big\|_2 \\ &= \big\|r - \tilde{r} + \Pi_{\tilde{A}}(\tilde{r} - r) + (p^t - r) - \Pi_{\tilde{A}}(p^t - r)\big\|_2 \\ &\le 2 \cdot \|r - \tilde{r}\|_2 + \cdot\|\Pi_A - \Pi_{\tilde{A}}\|_{\mathrm{op}} \cdot \|p^t - r\|_2 \\ &\le \frac{\delta}{12\sqrt{n}} + \frac{8\sqrt{2\delta}}{w_{\min}\zeta\sqrt{n}}.\end{aligned}$$

Also $\|\bar{p}^t - \hat{p}^t\|_2 \le \Big\|\sum_{j=1}^{k'}(b_j^\dagger p^t - \alpha^j_{\sigma^j(t)})b_j\Big\|_2 = O\Big(\frac{\sqrt{k\delta}}{w_{\min}^{1.5}\zeta\sqrt{n}}\Big)$

where the last equality follows from Lemma 4.7. Thus, $\|p^t - \tilde{p}^t\|_1 = O\big(\frac{\sqrt{k\delta}}{w_{\min}^{1.5}\zeta}\big)$. Also, we have $|w_t - \tilde{w}_t| = O\big(\frac{\delta\omega^{1.5}k^5}{w_{\min}^2\zeta^2}\big)$ by Lemma 4.7. Finally, note that $\mathrm{Tran}(w, P; \tilde{w}, \tilde{P})$ is at most

$$\begin{aligned}&\frac{1}{2}\Big(\sum_{t=1}^{k}\min\{w_t, \tilde{w}_t\}\max_t\|p^t - \tilde{p}^t\|_1 + \|w - \tilde{w}\|_1\max_{t\neq t'}\|p^t - \tilde{p}^{t'}\|_1\Big) \\ &\le \frac{1}{2}\Big(\sum_{t=1}^{k}\min\{w_t, \tilde{w}_t\}\max_t\|p^t - \tilde{p}^t\|_1 \\ &\qquad + \|w - \tilde{w}\|_1\big(\max_{t\neq t'}\|p^t - p^{t'}\|_1 + \max_t\|p^t - \tilde{p}^t\|_1\big)\Big) \\ &\le \max_t\|p^t - \tilde{p}^t\|_1 + \|w - \tilde{w}\|_1 \cdot \frac{2}{w_{\min}} = O\Big(\frac{\sqrt{k\delta}}{w_{\min}^{1.5}\zeta}\Big).\end{aligned}$$

The running time is dominated by the time required to compute the spectral decomposition in step A1.3, the calls to Learn in steps A2.2 and A3.2, and the time to compute $\tilde{p}^t$ in step A3.4. The other steps are clearly polytime. As noted in Remark 1, it suffices to compute a decomposition such that $\|\tilde{M} - \tilde{R} - \sum_{i=1}^n \lambda_i v_i v_i^\dagger\| = O\big(\frac{\delta}{n}\big)$; this takes time $\mathrm{poly}\big(n, \ln(n/\delta)\big)$. The LP used in step A3.4 is of polynomial size, and hence can be solved in polytime. Procedure Learn requires solving (Q_v); again, an approximate solution suffices and can be computed in polytime. Theorem 5.1 proves that the one-dimensional learning problem can be solved in polytime; hence, Learn takes polytime. $\square$

5. THE ONE-DIMENSIONAL PROBLEM: LEARNING MIXTURE SOURCES ON [0,1]

In this section, we supply the key subroutine called upon in step L2 of Algorithm Learn, which will complete the description of Algorithm 1. We are given a k-mixture source $\big(w, \pi_x(P)\big)$ on $\big[-\frac{1}{2}, \frac{1}{2}\big]$. (Recall that Learn invokes the procedure for the mixture $\big(w, \pi_{a/2H}(P)\big)$ where $\|a\|_\infty \le H$.) It is clear that we *cannot* in general reconstruct this mixture source with an aperture size that is independent of n, let alone aperture $2k-1$. However, our goal is somewhat different and more modest. We seek to reconstruct the k-spike distribution $\big(w, \mathrm{E}[\pi_x(P)]\big)$, and we show that this *can* be achieved with aperture $2k-1$ (which is the smallest aperture at which this is information-theoretically possible).

It is easy to obtain a $(2k-1)$-snapshot from $(w, \pi_x(P))$ given a $(2k-1)$-snapshot from (w, P) by simply replacing each item $i \in [n]$ that appears in the snapshot by x_i. We will assume in the sequel that every constituent $\pi_x(p^t)$ is supported on $[0,1]$, which is simply a translation by $\frac{1}{2}$.

To simplify notation, we use $\boldsymbol{\theta} = \big(\vartheta, (q^1, \ldots, q^k)\big)$ to denote the k-mixture source on $[0,1]$, and $\big(\vartheta, \alpha = (\alpha_1, \ldots, \alpha_k)\big)$ to denote the corresponding k-spike distribution, where $\alpha_i \in [0,1]$ is the expectation of q^i for all $i \in [k]$. We equivalently view (ϑ, α) as a k-mixture source $\big(\vartheta, (f^1, \ldots, f^k)\big)$ on $\{0,1\}$: each f^i is a "coin" whose bias is $f^i_1 = \alpha_i$. In Section 5.1, we describe how to learn such a *binary* mixture source from its $(2k-1)$-snapshots (see Algorithm 2 and Theorem 5.3). Thus, if we can obtain $(2k-1)$-snapshots from the binary source $\big(\vartheta, (f^1, \ldots, f^k)\big)$ (although our input is $\boldsymbol{\theta}$) then Theorem 5.3 would yield the the desired result. We show that this is indeed possible, and hence, obtain the following result (whose proof appears at the end of Section 5.1).

THEOREM 5.1. *Let $\boldsymbol{\theta} = \big(\vartheta, (q^1, \ldots, q^k)\big)$ be a k-mixture source on $[0,1]$, and (ϑ, α) be the corresponding k-spike distribution. Let $\tau = \min_{j \neq j'} |\alpha_j - \alpha_{j'}|$. For any $s < \tau$ and $\psi > 0$, using $3k2^{4k}s^{-4k}\ln(4k/\psi)$ $(2k-1)$-snapshots from source $\boldsymbol{\theta}$, one can compute in polytime a k-spike distribution $(\tilde{\vartheta}, \tilde{\alpha})$ on $[0,1]$ such that $\mathrm{Tran}(\vartheta, \alpha; \tilde{\vartheta}, \tilde{\alpha}) \leq 1024ks^{1/(4k)}$ with probability at least $1 - \psi$.*

5.1 Learning a binary k-mixture source

Recall that $\big(\vartheta, (f^1, \ldots, f^k)\big)$ denotes the binary k-mixture source, and $\alpha_i = f^i_1$ is the bias of the i-th "coin". We can collect from each $(2k-1)$-snapshot a random variable $0 \leq X \leq 2k-1$ denoting the number of times the outcome "1" occurs in the snapshot. Thus,

$$\Pr[X = i] = \binom{2k-1}{i} \sum_{j=1}^{k} \vartheta_j \alpha_j^i (1-\alpha_j)^{2k-1-i}.$$

Our objective is to use these statistics to reconstruct, in transportation distance (see Section 2.2), the binary source (i.e., the mixture weights and the k biases). Now consider the equivalent k-spike distribution (ϑ, α). The i-th moment, and (what we call) the i-th *normalized binomial moment* (NBM) of this distribution are respectively

$$g_i(\vartheta, \alpha) = \sum_{j=1}^{k} \vartheta_j \alpha_j^i, \qquad \nu_i(\vartheta, \alpha) = \sum_{j=1}^{k} \vartheta_j \alpha_j^i (1-\alpha_j)^{2k-1-i}$$

Up to the factors $\binom{2k-1}{i}$ the NBMs are precisely the statistics of the random variable X and so our objective in this section can be restated as: use the empirical NBMs to reconstruct the k-spike distribution (ϑ, α).

Let $g(\vartheta, \alpha) = \big(g_i(\vartheta, \alpha)\big)_{i=0}^{2k-1}$ and $\nu(\vartheta, \alpha) = \big(\nu_i(\vartheta, \alpha)\big)_{i=0}^{2k-1}$ denote the row-vectors of the first $2k-1$ moments and NBMs respectively of (ϑ, α). For an integer $b > 0$ and a vector $\beta = (\beta_1, \ldots, \beta_\ell)$, let $A_b(\beta)$ be the $\ell \times b$ matrix $(A_b(\beta))_{ij} = (1-\beta_i)^{b-1-j}\beta_i^j$ (with $1 \leq i \leq \ell$ and $0 \leq j \leq b-1$). Analogously, let $V_b(\beta)$ be the $\ell \times b$ "Vandermonde" matrix $(V_b(\beta))_{ij} = \beta_i^j$ (with $1 \leq i \leq \ell$ and $0 \leq j \leq b-1$). Let Pas be the $2k \times 2k$ lower-triangular "Pascal triangle" matrix with non-zero entries $\mathrm{Pas}_{ij} = \binom{2k-1-j}{i-j}$ for $0 \leq j \leq 2k-1$ and $j \leq i \leq 2k-1$. Then $V_{2k}(\alpha) = A_{2k}(\alpha)\mathrm{Pas}$, $\nu(\vartheta, \alpha) = \vartheta A_{2k}(\alpha)$, and $g(\vartheta, \alpha) = \vartheta V_{2k}(\alpha) = \nu(\vartheta, \alpha)\mathrm{Pas}$.

In our algorithm it is convenient to use the empirical ordinary moments, but what we obtain are actually the empirical NBMs, so we need the following lemma.

LEMMA 5.2. $\|\mathrm{Pas}\|_{\mathrm{op}} \leq 4^k/\sqrt{3}$.

PROOF. The non-zero entries in column j of Pas are $\binom{m}{\ell}$ for $\ell = 0, \ldots, m = 2k-1-j$. Therefore, $\|\mathrm{Pas}\|_{\mathrm{op}} \leq \|\mathrm{Pas}\|_F = \sqrt{\sum_{m=0}^{2k-1}\sum_{\ell=0}^{m}\binom{m}{\ell}^2}$. Since $\sum_{i=0}^{m}\binom{m}{i}^2 = \binom{2m}{m} \leq 2^{2m}$, we have $\|\mathrm{Pas}\|_F \leq \sqrt{\sum_{m=0}^{2k-1} 2^{2m}}$. $\square$

Our algorithm uses two input parameters τ and ξ as input, and the empirical NBM vector $\tilde{\nu}$, which we convert to an empirical moment vector $\tilde{g}$ by multiplying by Pas. Since we infer (in the sampling limit) the locations of the k spikes exactly, there is a singularity in the process when spikes coincide. So we assume a minimum separation between spikes: $\tau = \min_{j \neq j'} |\alpha_j - \alpha_{j'}|$. (It is of course possible to simply run a doubling search for sufficiently small τ, but the required accuracy in the moments, and hence sample size, does increase as τ decreases.) We also assume a bound ξ on the accuracy of our empirical statistics. (When we utilize Theorem 5.3 to obtain Theorem 5.1, ξ is a consequence, and not an input parameter). We require that

$$\|\tilde{\nu} - \nu(\vartheta, \alpha)\|_2 \leq \xi 4^{-k}\sqrt{3}, \qquad \xi \leq \tau^{2k} \tag{5}$$

THEOREM 5.3. *There is a polytime algorithm that receives as input τ, ξ, an empirical NBM vector $\tilde{\nu} \in \mathbb{R}^{2k}$ satisfying (5), and outputs a k-spike distribution $(\tilde{\vartheta}, \tilde{\alpha})$ on $[0,1]$ such that $\mathrm{Tran}(\vartheta, \alpha; \tilde{\vartheta}, \tilde{\alpha}) \leq O(\xi^{\Omega(1/k^2)})$.*

We first show the information-theoretic feasibility of Theorem 5.3: the transportation distance between two probability measures on $[0,1]$ is upper bounded by (a moderately-growing function of) the Euclidean distance between their moment maps. (To use Lemma 5.4 to prove Theorem 5.3, we have to also show how to compute $\tilde{\vartheta}$ and $\tilde{\alpha}$ from $\tilde{g}$ such that $\|\tilde{g} - g(\tilde{\vartheta}, \tilde{\alpha})\|_2$, and hence, $\|g(\vartheta, \alpha) - g(\tilde{\vartheta}, \tilde{\alpha})\|_2$ is small.)

LEMMA 5.4. *For any two (at most) k-spike distributions (ϑ, α) $(\tilde{\vartheta}, \tilde{\alpha})$ on $[0,1]$,*

$$\|g(\vartheta, \alpha) - g(\tilde{\vartheta}, \tilde{\alpha})\|_2 \geq \frac{1}{(2k-1)^{4k}2^{8k-5}} \cdot \Big(\mathrm{Tran}(\vartheta, \alpha; \tilde{\vartheta}, \tilde{\alpha})\Big)^{4k-2}.$$

Lemma 5.4 can be geometrically interpreted as follows. The point $g(\vartheta, \alpha)$ is in the convex hull of the moment curve and is therefore, by Caratheodory's theorem, expressible as a convex combination of $2k$ points on the curve. However, this point is special in that it belongs to the collection of points expressible as a convex combination of merely k points of the curve. Lemma 5.4 shows that $g(\vartheta, \alpha)$ is in fact *uniquely* expressible in this way, and that moreover this combination is stable: any nearby point in this collection can only be expressed as a very similar convex combination. We utilize the following lemma, which can be understood as a global curvature property of the moment curve; we defer its proof to the end of this section. We prove a partial converse of Lemma 5.4 in Section 6, and hence obtain a sample-size lower bound that is exponential in k. The moment curve plays a central role in convex and polyhedral geometry [8], but as far as we know Lemmas 5.4 and 5.5 are new, and may be of independent interest.

LEMMA 5.5. *Let $0 \leq \beta_1 < \ldots < \beta_{\kappa+1} \leq 1$, $\ell \in [\kappa]$, and $s = \beta_{\ell+1} - \beta_\ell$. Let $\gamma(x) = \sum_{i=0}^{\kappa}\gamma_i x^i$ be a real polynomial of degree κ evaluating to 1 at the points $\beta_1, \ldots, \beta_\ell$ and evaluating to 0 at the points $\beta_{\ell+1}, \ldots, \beta_{\kappa+1}$. Then $\sum_{i=0}^{\kappa}\gamma_i^2 \leq \kappa^2 2^{4\kappa-1}s^{-2\kappa}$.*

Proof of Lemma 5.4. Denote $\{\alpha_1,\ldots,\alpha_k\}\cup\{\tilde\alpha_1,\ldots,\tilde\alpha_k\}$ by $\overline{\alpha}=\{\overline{\alpha}_1,\ldots,\overline{\alpha}_K\}$ where $\overline{\alpha}_1<\ldots<\overline{\alpha}_K$. Define $\overline{\vartheta}_i=\sum_{j:\alpha_j=\overline{\alpha}_i}\vartheta_j-\sum_{j:\tilde\alpha_j=\overline{\alpha}_i}\tilde\vartheta_j$ for $i\in[K]$. Let $\overline{\vartheta}\in\mathbb{R}^K$ be the row vector $(\overline{\vartheta}_1,\ldots,\overline{\vartheta}_K)$. Let $\eta=\mathrm{Tran}(\vartheta,\alpha;\tilde\vartheta,\tilde\alpha)$. So we need to show that $\|\overline{\vartheta}V_{2k}(\overline{\alpha})\|_2\geq\frac{1}{(2k-1)^{4k}2^{8k-5}}\cdot\eta^{4k-2}$. It suffices to show that $\|\overline{\vartheta}V_K(\overline{\alpha})\|_2\geq\frac{1}{(K-1)^{2K}2^{4K-5}}\cdot\eta^{2K-2}$. There is an $1\leq\ell<K$ such that $\left|\sum_{i=1}^{\ell}\overline{\vartheta}_i\right|\cdot(\overline{\alpha}_{\ell+1}-\overline{\alpha}_\ell)\geq\eta/(K-1)$. Let $\delta=\sum_{i=1}^{\ell}\overline{\vartheta}_i$; without loss of generality $\delta\geq0$, and note that $\delta\leq1$. Let $s=\overline{\alpha}_{\ell+1}-\overline{\alpha}_\ell$, so $(K-1)\delta s\geq\eta$. Denote row i of a matrix Z by Z_{i*} and column j by Z_{*j}. We lower bound $\|\overline{\vartheta}V_K(\overline{\alpha})\|_2$, by considering its minimum value under the constraints $\sum_{i=1}^{\ell}\overline{\vartheta}_i=\delta$ and $\sum_{i=1}^{K}\overline{\vartheta}_i=0$.

A vector $y^\dagger=\overline{\vartheta}V_K(\overline{\alpha})$ minimizing $\|y\|_2$ must be orthogonal to $V_K(\overline{\alpha})_{i*}-V_K(\overline{\alpha})_{i'*}$ if $1\leq i<i'\leq\ell$ or if $\ell+1\leq i<i'\leq K$. This means that there are scalars c and d such that $V_K(\overline{\alpha})y=c(\sum_{j=1}^{\ell}e_j)+d(\sum_{j=\ell+1}^{K}e_j)$, where vector $e_j\in\mathbb{R}^K$ has a 1 in the j-th position and 0 everywhere else. Therefore, $y=c\gamma+d\gamma'$, where $\gamma=\sum_{j=1}^{\ell}(V_K(\overline{\alpha})^{-1})_{*j}$ and $\gamma'=\sum_{j=\ell+1}^{K}(V_K(\overline{\alpha})^{-1})_{*j}$. At the same time

$$\delta=\sum_{i=1}^{\ell}\overline{\vartheta}_i=\overline{\vartheta}V_K(\overline{\alpha})\gamma=y^\dagger\gamma=c\|\gamma\|_2^2+d\gamma'^\dagger\gamma$$

$$-\delta=\sum_{i=\ell+1}^{K}\overline{\vartheta}_i=\overline{\vartheta}V_K(\overline{\alpha})\gamma'=y^\dagger\gamma'=c\gamma^\dagger\gamma'+d\|\gamma'\|_2^2$$

and hence, $\|y\|_2^2=y\cdot(c\gamma+d\gamma')=(c-d)\delta$. Solving for c,d gives

$$c-d=\frac{\delta\|\gamma+\gamma'\|_2^2}{\|\gamma\|_2^2\cdot\|\gamma'\|_2^2-(\gamma^\dagger\cdot\gamma')^2}.$$

First we examine the numerator of $c-d$. Like any combination of the columns of $V_K(\overline{\alpha})^{-1}$, $\gamma+\gamma'$ is the list of coefficients of a polynomial of degree $K-1$, in the basis $1,x,\ldots,x^{K-1}$. By definition, $\gamma+\gamma'=\sum_j(V_K(\overline{\alpha})^{-1})_{*j}$, which is to say that for every i, $V_K(\overline{\alpha})_{i*}\cdot(\gamma+\gamma')=1$. So the polynomial $\gamma+\gamma'$ evaluates to 1 at every $\overline{\alpha}_i$. It can therefore only be the constant polynomial 1; this means that $(\gamma+\gamma')_i=1$ if $i=0$, and $(\gamma+\gamma')_i=0$ otherwise. Thus $\|\gamma+\gamma'\|_2^2=1$.

Next we examine the denominator, which we upper bound by $\|\gamma\|_2^2\cdot\|\gamma'\|_2^2$. When interpreted as a polynomial, γ takes the value 1 on a nonempty set of points $\overline{\alpha}_1,\ldots,\overline{\alpha}_\ell$ separated by the positive distance $s=\overline{\alpha}_{\ell+1}-\overline{\alpha}_\ell$ from another nonempty set of points $\overline{\alpha}_{\ell+1},\ldots,\overline{\alpha}_K$ upon which it takes the value 0. Observe that if the polynomial was required to change value by a large amount within a short interval, it would have to have large coefficients. A converse to this is the inequality stated in Lemma 5.5. Using this to bound $\|\gamma\|_2^2$ and $\|\gamma'\|_2^2$, and since $\delta s\geq\eta/(K-1)$, we obtain that

$$\|y\|_2^2=(c-d)\delta\geq\frac{\delta^2}{\|\gamma\|_2^2\cdot\|\gamma'\|_2^2}\geq\frac{\delta^2}{((K-1)^2 2^{4K-5}s^{-2K+2})^2}$$
$$\geq\frac{\eta^{4K-4}}{(K-1)^{4K}2^{8K-10}}.\quad\square$$

We now define the algorithm promised by Theorem 5.3. To give some intuition, suppose first that we are given the true moment vector $g(\vartheta,\alpha)=\vartheta V_{2k}(\alpha)$. Observe that there is a common vector $\lambda=(\lambda_0,\ldots,\lambda_k)^\dagger$ of length $k+1$ that is a dependency among every $k+1$ adjacent columns of $V_{2k}(\alpha)$. In other words, letting $\Lambda=\Lambda(\lambda)$ denote the $2k\times k$ matrix with $\Lambda_{ij}=\lambda_{i-j}$ (for $0\leq i<2k$, $0\leq j<k$ and with the understanding $\lambda_\ell=0$ for $\ell\notin\{0,\ldots,k\}$), $V_{2k}(\alpha)\Lambda=0$. Thus $g(\vartheta,\alpha)\Lambda=\vartheta V_{2k}(\alpha)\Lambda=0$. Overtly this is a system of $2k$ equations to determine λ. But we eliminate the redundancy in Λ by forming the $k\times(k+1)$ matrix $G=G(g(\vartheta,\alpha))$ defined by $G_{ij}=g(\vartheta,\alpha)_{i+j}$ for $i=0,\ldots,k-1$ and $j=0,\ldots,k$, and then solve the system of linear equations $G\lambda=0$ to obtain λ. This system does not have a unique solution, so in the sequel λ will denote a solution with $\lambda_k=1$. For each $i=1,\ldots,k$, we have $\left(V_{2k}(\alpha)\Lambda\right)_{i,1}=\sum_{\ell=0}^{k}\lambda_\ell{\alpha_i}^\ell=0$. This implies that we can obtain the α_i values by computing the roots of the polynomial $P_\lambda(x):=\sum_{\ell=0}^{k}\lambda_\ell x^\ell$. Once we have the α_i's, we can compute ϑ by solving the linear system $yV_{2k}(\alpha)=g(\vartheta,\alpha)$ for y.

Of course, we are actually given $\tilde g$ rather than the true vector $g(\vartheta,\alpha)$. So we need to control the error in estimating first α and then ϑ. The learning algorithm is as follows.

Algorithm 2. *Input: parameters ξ,τ and empirical moments $\tilde g$ such that $\|\tilde g-g(\vartheta,\alpha)\|_2\leq\xi$.*
Output: a k-spike distribution $(\tilde\vartheta,\tilde\alpha)$

B1. Solve the minimization problem:
$$\text{minimize}\quad\|x\|_1\quad\text{s.t.}\quad\|G(\tilde g)x\|_1\leq2^k k\xi,\quad x_k=1\qquad\text{(P)}$$
which can be encoded as a linear program and hence solved in polytime, to obtain a solution $\tilde\lambda$. Observe that since $G(\tilde g)$ has $k+1$ columns and k rows, there is always a feasible solution.

B2. Let $\bar\alpha_1,\ldots,\bar\alpha_k$ be the (possibly complex) roots of the polynomial $P_{\tilde\lambda}$. Thus, we have $V_{2k}(\bar\alpha)\Lambda(\tilde\lambda)=0$. We map the roots to values in $[0,1]$ as follows. Let $\epsilon=\frac{4}{\tau}(2k\xi)^{1/k}$. First we compute $\hat\alpha_1,\ldots,\hat\alpha_k$ such that $|\hat\alpha_i-\bar\alpha_i|\leq\epsilon$ for all i, in time $\mathrm{poly}\big(\log(\frac{1}{\epsilon})\big)$, using Pan's algorithm [38, Theorem 1.1][2]. We now set $\tilde\alpha_i=\max\{0,\min\{\mathrm{Re}(\hat\alpha_i),1\}\}$.

B3. Finally, we set $\tilde\vartheta$ to be the row-vector y that minimizes $\|yV_{2k}(\tilde\alpha)-\tilde g\|_2^2$ subject to $\|y\|_1=1$, $y\geq0$. Note that this is a convex quadratic program that can be solved exactly in polytime [17].

We now analyze Algorithm 2 and justify Theorem 5.3. Recall that $\tau=\min_{j\neq j'}|\alpha_j-\alpha_{j'}|$. We need the following lemma, whose proof appears at the end of this section.

Lemma 5.6. *The weights $\tilde\vartheta$ satisfy $\|\tilde\vartheta V_{2k}(\tilde\alpha)-\tilde g\|_2\leq\|g(\vartheta,\alpha)-\tilde g\|_2+\frac{(8k)^{5/2}}{\tau}\cdot(2k\xi)^{1/k}$.*

Proof of Theorem 5.3. We call Algorithm 2 with $\tilde g=\tilde\nu\mathrm{Pas}$. By Lemma 5.2, we obtain that $\|\tilde g-g(\vartheta,\alpha)\|_2\leq\xi$, and by Lemma 5.6, we have that $\|g(\vartheta,\alpha)-\tilde\vartheta V_{2k}(\tilde\alpha)\|_2\leq2\|g(\vartheta,\alpha)-\tilde g\|_2+\frac{8}{\tau}\cdot(8k)^{3/2}\big(2k\xi\big)^{1/k}$. Coupled with Lemma 5.4 and since $\xi\leq\tau^{2k}$, we obtain that $\mathrm{Tran}(\vartheta,\alpha;\tilde\vartheta,\tilde\alpha)$ is at most

$$\Big[(2k-1)^{4k}2^{8k-5}\|g(\vartheta,\alpha)-g(\tilde\vartheta,\tilde\alpha)\|_2\Big]^{\frac{1}{4k-2}}$$
$$\leq\left[(2k-1)^{4k}2^{8k-5}\Big(2\xi+\frac{(8k)^{\frac52}}{\tau}\big(2k\xi\big)^{\frac1k}\Big)\right]^{\frac{1}{4k-2}}$$
$$\leq\Big[(2k-1)^{4k}2^{8k-5}\Big(2\xi+(8k)^{5/2}\big(2k\sqrt{\xi}\big)^{1/2k}\Big)\Big]^{\frac{1}{4k-2}}$$
$$\leq1024\cdot k\xi^{\frac{1}{8k^2}}.\qquad\square$$

[2]The theorem requires that the complex roots lie within the unit circle and that the coefficient of the highest-degree term is 1; but the discussion following it in [38] shows that this is essentially without loss of generality.

Proof of Theorem 5.1. We convert $\boldsymbol{\theta}$ to the corresponding binary source $\big(\vartheta, (f^1, \ldots, f^k)\big)$ by randomized rounding. Given a $(2k-1)$-snapshot $z = (z_1, \ldots, z_{2k-1}) \in [0,1]^{2k-1}$ from $\boldsymbol{\theta}$, we obtain a $(2k-1)$-snapshot from the binary source as follows. We choose $2k-1$ independent values $a_1, \ldots, a_{2k-1}$ uniformly at random from $[0,1]$ and set $X_i = 1$ if $z_i \geq a_i$ and 0 otherwise for all $i \in [2k-1]$. Note that if q^j is the constituent generating the $(2k-1)$-snapshot z, then $\Pr[X_i = 1 | q^j] = \mathrm{E}[X_i | q^j] = \alpha_j$, and so $X_1, \ldots, X_{2k-1}$ is a random $(2k-1)$-snapshot from the above binary source.

Now we apply Theorem 5.3, setting $\xi = s^{2k}$. Let $\tilde{\nu}$ be the empirical NBM-vector obtained from the $(2k-1)$-snapshots of the above binary source (i.e., $\tilde{\nu}_i = \binom{2k-1}{i}^{-1}$ (frequency with which the $(2k-1)$-snapshot has exactly i 1s)). The stated sample size ensures, via a Chernoff bound, that $\Pr\big[|\tilde{\nu}_i - \nu(\vartheta,\alpha)_i| \geq \frac{\xi 4^{-k}}{\sqrt{6k}}\big] < \frac{\psi}{2k}$ for all $i = 0, \ldots, 2k-1$. Hence, with probability at least $1-\psi$, we have $\|\tilde{\nu} - \nu(\vartheta,\alpha)\|_2 \leq \sqrt{2k} \cdot \|\tilde{\nu} - \nu(\vartheta,\alpha)\|_\infty \leq \xi 4^{-k}/\sqrt{3}$. $\square$

Proof of Lemma 5.5. There are two easy cases to dismiss before we reach the more subtle part of this lemma. The first easy case is $\ell = 1$. In this case γ is a single Lagrange interpolant: $\gamma(x) = \prod_{j=2}^{\kappa+1} \frac{x-\beta_j}{\beta_1-\beta_j}$. For $0 \leq i \leq \kappa$ let $e_i^\kappa(\beta_2, \ldots, \beta_{\kappa+1})$ be the i'th *elementary symmetric mean*,

$$e_i^\kappa(\beta_2, \ldots, \beta_{\kappa+1}) = \frac{1}{\binom{\kappa}{i}} \sum_{S \subseteq \{2,\ldots,\kappa+1\} : |S| = i} \prod_{j \in S} \beta_j$$

and observe that for all i, $0 \leq e_i^\kappa(\beta_2, \ldots, \beta_{\kappa+1}) \leq 1$. Now

$$\gamma(x) = \Big(\prod_{j=2}^{\kappa+1} \frac{1}{\beta_1 - \beta_j}\Big) \sum_{i=0}^{\kappa} (-1)^{\kappa-i} \binom{\kappa}{i} e_{\kappa-i}^\kappa(\beta_2, \ldots, \beta_{\kappa+1}) x^i$$

So $\sum_{i=0}^\kappa \gamma_i^2 = \Big(\prod_{j=2}^{\kappa+1} \frac{1}{\beta_1-\beta_j}\Big)^2 \sum_{i=0}^\kappa \big(\binom{\kappa}{i} e_i^\kappa(\beta_2, \ldots, \beta_{\kappa+1})\big)^2 \leq s^{-2\kappa} \sum_{i=0}^\kappa \binom{\kappa}{i}^2 = \binom{2\kappa}{\kappa} s^{-2\kappa}$.

The second easy case is $\ell = \kappa$; this is almost as simple. Merely note that the above argument applies to the polynomial $1-\gamma$, so that we have only to allow for the possible increase of $|\gamma_0|$ by 1. Hence $\sum_{i=0}^\kappa \gamma_i^2 \leq 4\binom{2\kappa}{\kappa} s^{-2\kappa}$.

We now consider the less trivial case of $1 < \ell < \kappa$. The difficulty here is that the Lagrange interpolants of γ may have very large coefficients, particularly if among $\beta_1, \ldots, \beta_\ell$ or among $\beta_{\ell+1}, \ldots, \beta_{\kappa+1}$ there are closely spaced roots, as well there may be. We must show that these large coefficients cancel out in γ.

The trick is to examine not γ but $\partial\gamma/\partial x$. The roots of the derivative interlace the two sets on which γ is constant, which is to say, with $\beta_1' \leq \ldots \leq \beta_{\kappa-1}'$ denoting the roots of $\partial\gamma/\partial x$, that for $j < \ell$, $\beta_j \leq \beta_j' \leq \beta_{j+1}$, and for $j \geq \ell$, $\beta_{j+1} \leq \beta_j' \leq \beta_{j+2}$. In particular, none of the roots fall in the interval $(\beta_\ell, \beta_{\ell+1})$. For some constant C we can write $\partial\gamma/\partial x = C \prod_{j=0}^{\kappa-1}(x - \beta_j')$ (with $\mathrm{sign}(C) = (-1)^{1+\kappa-\ell}$). Observe that $\int_{\beta_\ell}^{\beta_{\ell+1}} \frac{\partial\gamma}{\partial x}(x)\, dx = -1$. So $(-1)^{1+\kappa-\ell}/C = \int_{\beta_\ell}^{\beta_{\ell+1}} (-1)^{\kappa-\ell} \prod_{j=0}^{\kappa-1}(x-\beta_j')\, dx$. Observe that if for any $j < \ell$, β_j' is increased, or if for any $j \geq \ell$, β_j' is decreased, then the integral decreases. So $(-1)^{1+\kappa-\ell}/C \geq \int_{\beta_\ell}^{\beta_{\ell+1}} (-1)^{\kappa-\ell}(x-\beta_\ell)^{\ell-1}(x-\beta_{\ell+1})^{\kappa-\ell}\, dx$. This is a definite integral that can be evaluated in closed form:

$$\int_{\beta_\ell}^{\beta_{\ell+1}} (-1)^{\kappa-\ell}(x-\beta_\ell)^{\ell-1}(x-\beta_{\ell+1})^{\kappa-\ell}\, dx = (\beta_{\ell+1}-\beta_\ell)^\kappa (\ell-1)!(\kappa-\ell)!/\kappa!\ .$$

Hence, $(-1)^{1+\kappa-\ell} C \leq \frac{\kappa!}{s^\kappa(\ell-1)!(\kappa-\ell)!}$. The sum of squares of coefficients of $\frac{\partial\gamma}{\partial x}$ is $C^2 \sum_{i=0}^{\kappa-1} \binom{\kappa-1}{i}^2 (e_i^{\kappa-1}(\beta_1', \ldots, \beta_{\kappa-1}'))^2 \leq C^2 \binom{2\kappa-2}{\kappa-1}$. Integration only decreases the magnitude of the coefficients, so the same bound applies to γ, with the exception of the constant coefficient. The constant coefficient can be bounded by the fact that γ has a root in $(0,1)$, and that in that interval the derivative is bounded in magnitude by $C \sum_{i=0}^{\kappa-1} \binom{\kappa-1}{i} = C \cdot 2^\kappa$. So $|\gamma_0| \leq C \cdot 2^\kappa$. Consequently, $\sum_{i=0}^\kappa \gamma_i^2$ is at most

$$C^2 \left[\binom{2\kappa-2}{\kappa-1} + 2^{2\kappa}\right] \leq \frac{\binom{2\kappa-2}{\kappa-1} + 2^{2\kappa}}{s^{2\kappa}} \cdot \left(\frac{\kappa!}{(\ell-1)!(\kappa-\ell)!}\right)^2 \leq \frac{5\kappa^2 2^{2\kappa-2}}{s^{2\kappa}} \cdot \binom{\kappa-1}{\ell-1}^2 \leq \frac{5\kappa^2 2^{4\kappa-4}}{s^{2\kappa}},$$

which completes the proof of the lemma. $\square$

Proof of Lemma 5.6.

Recall that $G = G(g(\vartheta,\alpha))$ is the $k \times (k+1)$ matrix defined by $G_{ij} = g(\vartheta,\alpha)_{i+j}$ for $i = 0, \ldots, k-1$ and $j = 0, \ldots, k$; λ is such that $G\lambda = 0$ and $\lambda_k = 1$; $\Lambda = \Lambda(\lambda)$ is the $2k \times k$ matrix with $\Lambda_{ij} = \lambda_{i-j}$ (for $0 \leq i < 2k$, $0 \leq j < k$ with the understanding $\lambda_\ell = 0$ for $\ell \notin \{0, \ldots, k\}$; and $P_\lambda(x)$ is the polynomial $\sum_{\ell=0}^k \lambda_\ell x^\ell$. We use V_k, V_{2k}, to denote $V_k(\alpha), V_{2k}(\alpha)$ respectively, and $\tilde{V}_k, \tilde{V}_{2k}, \tilde{G}, \tilde{\Lambda}$ to denote $V_k(\tilde{\alpha}), V_{2k}(\tilde{\alpha}), G(\tilde{g}), \Lambda(\tilde{\lambda})$ respectively. We abbreviate $g(\vartheta,\alpha)$ to g.

Lemma 5.7. *If $\|\tilde{g} - g\|_2 \leq \xi$, then $\|G\tilde{\lambda}\|_1 \leq 2^{k+1} k\xi$.*

Proof. First, observe that $\tilde{G}\lambda = G\lambda + (\tilde{G} - G)\lambda = (\tilde{G} - G)\lambda$. Also $\|\lambda\|_2 \leq \|\lambda\|_1 = \prod_{i=1}^k (1+\alpha_i) \leq 2^k$. The last two inequalities follows since $P_\lambda(x) = \prod_{i=1}^k (x - \alpha_i)$, and $P_\lambda(-1) = (-1)^k \|\lambda\|_1$. So for any $i = 1, \ldots, k$, $\big|(G - \tilde{G})_i \cdot \lambda\big| \leq \|\lambda\|_2 \|G_i - \tilde{G}_i\|_2 \leq 2^k \xi$. Thus, λ is a feasible solution to (P), which implies that $\|\tilde{\lambda}\|_1 \leq 2^k$. We have $\|G\tilde{\lambda}\|_1 \leq \|\tilde{G}\tilde{\lambda}\|_1 + \|(G - \tilde{G})\tilde{\lambda}\|_1 \leq 2^k k\xi + \|(G - \tilde{G})\tilde{\lambda}\|_1$. For any $i = 1, \ldots, k$, $\big|(G - \tilde{G})_i \cdot \tilde{\lambda}\big| \leq \|G_i - \tilde{G}_i\|_2 \|\tilde{\lambda}\|_2 \leq 2^k \xi$, so $\|G\tilde{\lambda}\|_1 \leq 2^{k+1} k\xi$. $\square$

Lemma 5.8. *For every α_i, $i = 1, \ldots, k$, there exists a $\sigma(i) \in \{1, \ldots, k\}$ such that $\vartheta_i |\alpha_i - \tilde{\alpha}_{\sigma(i)}| \leq \frac{8}{\tau}(2k\xi)^{1/k}$.*

Proof. Since $\|G\tilde{\lambda}\|_2 \leq 2^{k+1} k\xi$ (by Lemma 5.7), we have equivalently that the $\|.\|_2$ norm of $g\tilde{\Lambda} = \vartheta V_{2k} \tilde{\Lambda}$ is at most $2^{k+1} k\xi$. We may write $\vartheta V_{2k} \tilde{\Lambda}$ as

$$\vartheta V_{2k} \tilde{\Lambda} = \vartheta \begin{pmatrix} P_{\tilde{\lambda}}(\alpha_1) & \alpha_1 P_{\tilde{\lambda}}(\alpha_1) & \cdots & \alpha_1^{k-1} P_{\tilde{\lambda}}(\alpha_1) \\ P_{\tilde{\lambda}}(\alpha_2) & \alpha_2 P_{\tilde{\lambda}}(\alpha_2) & \cdots & \alpha_2^{k-1} P_{\tilde{\lambda}}(\alpha_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_{\tilde{\lambda}}(\alpha_k) & \alpha_k P_{\tilde{\lambda}}(\alpha_k) & \cdots & \alpha_k^{k-1} P_{\tilde{\lambda}}(\alpha_k) \end{pmatrix}$$

which is equal to $\vartheta' V_k(\alpha)$ where $\vartheta' = \big(\vartheta_1 P_{\tilde{\lambda}}(\alpha_1), \cdots, \vartheta_k P_{\tilde{\lambda}}(\alpha_k)\big)$. Thus, we are given that $\|\vartheta' V_k\|_2 \leq 2^{k+1} k\xi$.

Let $(\gamma^i)^\dagger = \big(\arg\min_{y\in\mathbb{R}^k: y_i=1} \|yV_k\|_2\big)V_k$. Then, we also have $\|\vartheta' V_k\|_2 \geq \max_i |\vartheta'_i|\|\gamma^i\|_2$. Note that γ^i must be orthogonal to $(V_k)_{j*}$ for all $j \neq i$, and $(V_k)_{i*}\gamma^i = \|\gamma^i\|_2^2$. (Recall that Z_{i*} denotes row i of a matrix Z.) Let $Q_i(x) = \sum_{\ell=0}^{k-1} \gamma^i_\ell x^\ell$. Then, $Q_i(x) = \|\gamma^i\|_2^2 \prod_{j\neq i} \frac{x-\alpha_j}{\alpha_i-\alpha_j}$. Also, since the coefficients of $Q_i(x)$ have alternating signs, we have

$$|Q_i(-1)| = \|\gamma^i\|_1 = \|\gamma^i\|_2^2 \prod_{j\neq i} \frac{1+\alpha_j}{|\alpha_i-\alpha_j|}.$$

Hence, $\|\gamma^i\|_2 \geq \prod_{j\neq i} \frac{|\alpha_i-\alpha_j|}{1+\alpha_j}$. So we obtain the lower bound

$$\begin{aligned}\|\vartheta' V_k\|_2 &\geq \max_i \Big(|\vartheta'_i| \cdot \prod_{j\neq i} \frac{|\alpha_i-\alpha_j|}{1+\alpha_j}\Big)\\ &\geq \max_i \left(\vartheta_i \Big(\frac{\tau}{2}\Big)^{k-1} \prod_{j=1}^{k} |\alpha_i - \bar{\alpha}_j|\right)\\ &\geq \max_i \left(\vartheta_i \Big(\frac{\tau}{2}\Big)^{k-1} \prod_{j=1}^{k} |\alpha_i - \mathrm{Re}(\bar{\alpha}_j)|\right).\end{aligned}$$

The last inequality follows since complex roots occur in conjugate pairs, so if $\bar{\alpha}_\ell = a + bi$ is complex, then there must be some ℓ' such that $\bar{\alpha}_{\ell'} = a - bi$ and therefore,

$$\begin{aligned}\prod_j |\alpha_i - \bar{\alpha}_j| &= \big((\alpha_i - a)^2 + b^2\big) \cdot \prod_{j\neq \ell,\ell'} |\alpha_i - \bar{\alpha}_j|\\ &\geq (\alpha_i - a)^2 \cdot \prod_{j\neq \ell,\ell'} |\alpha_i - \bar{\alpha}_j|.\end{aligned}$$

Now, we claim that $|\alpha_i - \mathrm{Re}(\bar{\alpha}_j)| \geq \big||\alpha_i - \tilde{\alpha}_j| - \epsilon\big|$ for every j. If both $\mathrm{Re}(\bar{\alpha}_j)$ and $\mathrm{Re}(\hat{\alpha}_j)$ lie in $[0,1]$, or both of them are less than 0, or both are greater than 1, then this follows since $|\bar{\alpha}_j - \hat{\alpha}_j| \leq \epsilon$ and $\alpha_i \in [0,1]$. If $\mathrm{Re}(\bar{\alpha}_j) \notin [0,1]$ but $\mathrm{Re}(\hat{\alpha}_j) \in [0,1]$, or if $\mathrm{Re}(\bar{\alpha}_j) \in [0,1]$ but $\mathrm{Re}(\hat{\alpha}_j) \notin [0,1]$, then this again follows since $|\bar{\alpha}_j - \hat{\alpha}_j| \leq \epsilon$. Combining everything, we get that

$$2^k(2k\xi) \;\geq\; \|\vartheta' V_k\|_2 \;\geq\; \max_i \left(\vartheta_i \Big(\frac{\tau}{2}\Big)^{k-1} \prod_{j=1}^{k} \big||\alpha_i - \tilde{\alpha}_j| - \epsilon\big|\right).$$

This implies that for every $i = 1,\ldots,k$, there exists $\sigma(i) \in \{1,\ldots,k\}$ such that $\vartheta_i|\alpha_i - \tilde{\alpha}_{\sigma(i)}| \leq \frac{4}{\tau} \cdot \big(2k\xi\big)^{1/k} + \epsilon$. $\square$

We can now wrap up the proof of Lemma 5.6. Let $\eta = \frac{8}{\tau} \cdot \big(2k\xi\big)^{1/k}$. We will bound $\|\tilde{\vartheta}\tilde{V}_{2k} - \tilde{g}\|_2$ by exhibiting a solution $y \in [0,1]^k$, $\|y\|_1 = 1$ such that $\|y\tilde{V}_{2k} - \tilde{g}\|_2 \leq \|g - \tilde{g}\| + k(8k)^{3/2}\eta$. Let σ be the function whose existence is proved in Lemma 5.8. For $j = 1,\ldots,k$, set $y_j = \sum_{i:\sigma(i)=j} \vartheta_i$ (if $\sigma^{-1}(j) = \emptyset$, then $y_j = 0$). We have $\|y\tilde{V}_{2k} - \tilde{g}\|_2 \leq \|g - \tilde{g}\|_2 + \|g - y\tilde{V}_{2k}\|_2$. We expand $g - y\tilde{V}_{2k} = \vartheta V_{2k} - y\tilde{V}_{2k} = \sum_{i=1}^k \vartheta_i\big((V_{2k})_{i*} - (\tilde{V}_{2k})_{\sigma(i)*}\big)$ For every i,

$$\vartheta_i^2\|(V_{2k})_{i*} - (\tilde{V}_{2k})_{\sigma(i)*}\|_2^2 = \vartheta_i^2 \sum_{\ell=0}^{2k-1} (\alpha_i^\ell - \tilde{\alpha}_{\sigma(i)}^\ell)^2 \leq \vartheta_i^2 \cdot 8k^3 \cdot \eta^2.$$

Therefore, $\|g - y\tilde{V}_{2k}\|_2 \leq k(8k)^{3/2}\eta$. $\square$

6. LOWER BOUNDS

In this section, we prove sample-size and aperture lower bounds that apply even to the setting where we have k-mixture sources on $\{0,1\}$ (so $n = 2$). Recall that a k-mixture source on $\{0,1\}$ may be equivalently viewed as a k-spike distribution supported on $[0,1]$; in the sequel, we therefore focus on k-spike distributions. The separation of a k-spike distribution (or the equivalent k-mixture source) is the minimum separation between its spikes. Theorem 6.2 proves that $2k-1$ is the smallest aperture at which it becomes possible to reconstruct a k-spike distribution. We emphasize that this is an *information-theoretic* lower bound. We show (Theorem 6.2) that there are two k-spike distributions supported on $[0,1]$ having separation $\Omega\big(\frac{1}{k}\big)$ and transportation distance $\Omega\big(\frac{1}{k}\big)$ that yield exactly the same first $2k-2$ moments. Moreover, for any $b \geq 2k-1$, by adjusting the constant in the $\Omega(.)$s, one can ensure that the $(2k-1)$-th, $\ldots$, b-th moments of these two k-spike distributions are exponentially close.

It follows immediately that even with infinite sample size it is impossible to reconstruct a k-mixture source (with arbitrarily small error) if we limit the aperture to $2k-2$. Furthermore, we leverage the exponential closeness of the moments to show that for any aperture $b \geq 2k-1$, there exists $\tau = \Omega\big(\frac{1}{k}\big)$ such that reconstructing a k-mixture source on $\{0,1\}$ having separation τ to within transportation distance $\frac{\tau}{4}$ requires *exponential in* k sample size (Theorem 6.1). In fact, since $n = 2$, this means that with arbitrary mixtures, the exponential dependence of the sample size on k remains *even with aperture* $O(k\log n)$, and more generally, even with aperture $O\big(k \cdot \kappa(n)\big)$ for any function $\kappa(.)$. (To place this in perspective, observe that with separation $\tau = \Omega\big(\frac{1}{k}\big)$, if we have $\Omega(k^2\log k)$ aperture, then $O(k^3)$ samples suffice to reconstruct the given mixture source to within transportation distance $\frac{\tau}{4}$. This is because with, with high probability, we will see every $\{0,1\}$ source or "coin" with weight $\vartheta_i \geq \frac{1}{\tau^2}$, and we can estimate its bias within additive error, say $\frac{\tau}{8}$, with probability $1 - \frac{1}{\mathrm{poly}(k)}$ since we have $\Omega\big(\frac{\log k}{\tau^2}\big)$ coin tosses available. The unseen coins contribute $O(\tau)$ to the transportation distance, so we infer k-mixture source within transportation distance $\frac{\tau}{4}$.)

THEOREM 6.1. *(i) With aperture $2k-2$, it is impossible to reconstruct a k-mixture source having separation at least $\frac{1}{2k}$ to within transportation distance $\frac{1}{8k}$ even with infinite sample size.*
(ii) For any $\psi \in (0,1)$, and any constants $c_A \geq 1$, $c_E \geq 0$, there exists $\tau = \Omega\big(\frac{1}{k}\big)$ such that reconstructing a k-mixture source having separation τ to within transportation distance $\frac{\tau}{4}$ with probability at least $1-\psi$ using aperture $c_A(2k-1)$ requires $\Omega\big(3^{c_E k}\ln(\frac{1}{\psi})\big)$ samples.

Our approach for proving Theorem 6.1 is as follows. To prove the existence of two suitable k-spike distributions (Theorem 6.2), we fix some spike locations ensuring the required separation and transportation-distance bounds, and search for suitable probability weights to place on these locations so as to obtain the desired closeness of moments for the two k-spike distributions. Since moments are linear functions of the weights (and the spike locations are fixed), this search problem can be encoded as a minimization LP (P1). To upper bound the optimum, we move to the dual LP (D1), which can be interpreted as finding a polynomial satisfying certain conditions on its coefficients and roots, to maximize the variation between its values at certain spike locations. We upper bound the variation possible by such a polynomial using elementary properties of polynomials. Finally,

the closeness of moments of the two k-spike distributions obtained this way implies that the distributions of b-snapshots of these two distributions have exponentially small variation distance (Lemma 6.3), and this yields the sample-size lower bound in Theorem 6.1.

THEOREM 6.2. *Let $b \geq 2k-1$, and $\rho \geq 2$. There are two k-spike distributions (y,α) and (z,β) on $[0,1]$ with separation $\frac{2}{(2k-1)\rho}$ and $\mathrm{Tran}(y,\alpha;z,\beta) \geq \frac{1}{(2k-1)\rho}$ such that $g_\ell(y,\alpha) = g_\ell(z,\beta)$ for all $\ell = 0,\ldots,2k-2$, and $\sum_{\ell=2k-1}^{b}\binom{b}{\ell}2^\ell|g_\ell(y,\alpha) - g_\ell(z,\beta)| \leq 4\cdot\frac{3^b}{\rho^{2k-1}}$.*

PROOF. Let $\epsilon = \frac{1}{\rho}$. We set $\alpha_i = \epsilon\cdot\frac{2(i-1)}{2k-1}$, and $\beta_i = \epsilon\cdot\frac{2i-1}{2k-1} = \alpha_i + \frac{\epsilon}{2k-1}$ for all $i = 1,\ldots,k$. Note that for any mixture weights $y_1,\ldots,y_k$, and $z_1,\ldots,z_k$, the separation of (y,α) and (z,β) is $\frac{2}{(2k-1)\rho}$ and $\mathrm{Tran}(y,\alpha;z,\beta) \geq \frac{1}{(2k-1)\rho}$. We obtain y and z by solving the following linear program (P1), whose optimal value we show is at most $4\cdot\frac{3^b}{\rho^{2k-1}}$.

$$\min \quad \sum_{\ell=2k-1}^{b}\binom{b}{\ell}2^\ell\lambda_\ell \tag{P1}$$

$$\text{s.t.} \quad \sum_{i=1}^{k}\big(z_i\beta_i^\ell - y_i\alpha_i^\ell\big) = 0 \qquad \forall \ell = 0,\ldots,2k-2 \tag{6}$$

$$\sum_{i=1}^{k}\big(z_i\beta_i^\ell - y_i\alpha_i^\ell\big) \leq \lambda_\ell \qquad \forall \ell = 2k-1,\ldots,b \tag{7}$$

$$\sum_{i=1}^{k}\big(y_i\alpha_i^\ell - z_i\beta_i^\ell\big) \leq \lambda_\ell \qquad \forall \ell = 2k-1,\ldots,b \tag{8}$$

$$\sum_{i=1}^{k} y_i = 1, \quad y_i, z_i \geq 0 \qquad \forall i = 1,\ldots,k.$$

The dual of (P1) is the following linear program.

$$\max \quad c \tag{D1}$$

$$\text{s.t.} \quad c - \sum_{\ell=0}^{2k-2}\gamma_\ell\alpha_i^\ell - \sum_{\ell=2k-1}^{b}(\gamma_\ell - \theta_\ell)\alpha_i^\ell \leq 0 \qquad \forall i = 1,\ldots,k$$

$$\sum_{\ell=0}^{2k-2}\gamma_\ell\beta_i^\ell + \sum_{\ell=2k-1}^{b}(\gamma_\ell - \theta_\ell)\beta_i^\ell \leq 0 \qquad \forall i = 1,\ldots,k$$

$$\gamma_\ell + \theta_\ell \leq \binom{b}{\ell}2^\ell \qquad \forall \ell = 2k-1,\ldots,b$$

$$\gamma_\ell, \theta_\ell \geq 0 \qquad \forall \ell.$$

The dual variable c corresponds to $\sum_i y_i = 1$, variables γ_ℓ for $\ell = 0,\ldots,b$ correspond to (6) and (8), and variables θ_ℓ for $\ell = 2k-1,\ldots,b$ correspond to (7). Given a feasible solution to (D1), if we set $\gamma'_\ell = \gamma_\ell - \min\{\gamma_\ell,\theta_\ell\}$, $\theta'_\ell = \theta_\ell - \min\{\gamma_\ell,\theta_\ell\}$ for all $\ell = 2k-1,\ldots,b$, then we obtain another feasible solution to (D1), where $\gamma'_\ell + \theta'_\ell = |\gamma'_\ell - \theta'_\ell| = |\gamma_\ell - \theta_\ell|$. Thus, an optimal solution to (D1) can be interpreted as a polynomial $f(x) = \sum_{\ell=0}^{b} f_\ell x^\ell$ satisfying $|f_\ell| \leq \binom{b}{\ell}2^\ell$ for all $\ell = 2k-1,\ldots,b$, and $f(\alpha_i) \geq c > 0 \geq f(\beta_i)$ for all $i = 1,\ldots,k$ (where $c > 0$ follows from Lemma 5.4).

Let $c' = 3^b\cdot\frac{\rho/(\rho-1)}{\rho^{2k-1}} \leq 2\cdot\frac{3^b}{\rho^{2k-1}}$. Suppose that $c > 2c'$. Observe that for any $x \in [0,\epsilon]$, by the Cauchy-Schwarz inequality and since the ℓ_2-norm is at most the ℓ_1 norm, we have

$$\begin{aligned}\Big|\sum_{\ell=2k-1}^{b} f_\ell x^\ell\Big| &\leq \Big(\sum_{\ell=2k-1}^{b}|f_\ell|\Big)\Big(\sum_{\ell=2k-1}^{b} x^\ell\Big) \\ &\leq \sum_{\ell=2k-1}^{b}\binom{b}{\ell}2^\ell\cdot\frac{\rho/(\rho-1)}{\rho^{2k-1}} \leq c'.\end{aligned} \tag{9}$$

Let $h(x) = \sum_{\ell=0}^{2k-2} f_\ell x^\ell - c/2$. Then, due to (9), we have $f(x) - c/2 - c' \leq h(x) \leq f(x) - c/2 + c'$ for all $x \in [0,\epsilon]$, so $h(\alpha_i) > 0 > h(\beta_i)$ for all $i = 1,\ldots,k$. But then $h(x)$ has $2k-1$ roots—one in every (α_i,β_i) and (β_i,α_{i+1}) interval—which is impossible since $h(x)$ is a polynomial of degree $2k-2$. □

Given a k-spike distribution $\big(\vartheta, \alpha = (\alpha_1,\ldots,\alpha_k)\big)$ on $[0,1]$, we abuse notation and denote the equivalent k-mixture source on $\{0,1\}$ also by (ϑ,α); that is, $\boldsymbol{\theta} = (\vartheta,\alpha)$ represents a mixture of k "coins", where coin i has bias α_i and is chosen with weight ϑ_i. Let $g(\vartheta,\alpha) = \big(g_i(\vartheta,\alpha)\big)_{i=0}^{2k-1}$. We use $D^{\boldsymbol{\theta}}$ (viewed as a vector in $\mathbb{R}_{\geq 0}^{\{0,1\}^{2k-1}}$) to denote the distribution of $(2k-1)$-snapshots induced by $\boldsymbol{\theta}$ on $\{0,1\}^{2k-1}$. The total variation distance $d_{\mathrm{TV}}(D^{\boldsymbol{y}}, D^{\boldsymbol{z}})$ between two such distributions is defined to be $\frac{1}{2}\|D^{\boldsymbol{y}} - D^{\boldsymbol{z}}\|_1$.

LEMMA 6.3. *Let $b \geq 2k-1$. Given two k-mixture sources $\boldsymbol{y} = (y,\alpha)$ and $\boldsymbol{z} = (z,\beta)$ on $\{0,1\}$ with identical first $2k-2$ moments, we have $d_{TV}(D^{\boldsymbol{y},b}, D^{\boldsymbol{z},b}) = \frac{1}{2}\sum_{\ell=2k-1}^{b}\binom{b}{\ell}2^\ell|g_\ell(y,\alpha) - g_\ell(z,\beta)|$.*

PROOF. For any $s \in \{0,1\}^b$ with i 1s, we have $D_s^{\boldsymbol{y},b} = \nu_i(y,\alpha)$ and $D_s^{\boldsymbol{z},b} = \nu_i(z,\beta)$. Therefore, $d_{\mathrm{TV}}(D^{\boldsymbol{y},b}, D^{\boldsymbol{z},b}) = \frac{1}{2}\sum_{i=0}^{b}\binom{b}{i}|\nu_i(y,\alpha) - \nu_i(z,\beta)|$. Let B be the $(b+1)\times(b+1)$ diagonal matrix with $B_{ii} = \binom{b}{i}$ for $0 \leq i \leq b$. Then, $\binom{b}{i}\nu_i(y,\alpha) = \big(g^b(y,\alpha)\mathrm{Pas}_{b+1}^{-1}B\big)_i$. Let $\Delta g^b := g^b(y,\alpha) - g^b(z,\beta)$. So $d_{\mathrm{TV}}(D^{\boldsymbol{y},b}, D^{\boldsymbol{z},b}) = \frac{1}{2}\|(\Delta g^b)(\mathrm{Pas}_{b+1}^{-1}B)\|_1$. We prove below that Pas_{b+1}^{-1} is the lower triangular matrix with entries $(\mathrm{Pas}_{b+1}^{-1})_{ij} = (-1)^{i-j}\binom{b-j}{i-j}$ for $0 \leq j \leq i \leq b$ (and 0 otherwise). Let Z_{i*} denote row i of matrix Z. Since $(\Delta g^b)_i = 0$ for $i = 0,\ldots,2k-2$, we have that $d_{\mathrm{TV}}(D^{\boldsymbol{y},b}, D^{\boldsymbol{z},v})$ is at most

$$\begin{aligned}\frac{1}{2}\sum_{\ell=2k-1}^{b}|(\Delta g^b)_\ell|\cdot\|(\mathrm{Pas}_{b+1}^{-1}B)_{\ell,*}\|_1 &= \frac{1}{2}\sum_{\ell=2k-1}^{b}|(\Delta g^b)_\ell|\sum_{j=0}^{\ell}\binom{b-j}{\ell-j}\binom{b}{j} \\ &= \frac{1}{2}\sum_{\ell=2k-1}^{b}|(\Delta g^b)_\ell|\binom{b}{\ell}\sum_{j=0}^{\ell}\binom{\ell}{j} \\ &= \frac{1}{2}\sum_{\ell=2k-1}^{b}|(\Delta g^b)_\ell|\binom{b}{\ell}2^\ell.\end{aligned}$$

To see the claim about Pas_{b+1}^{-1}, let Q be the claimed inverse matrix; so $Q_{ij} = (-1)^{i-j}\binom{b-j}{i-j}$ for $0 \leq j \leq i \leq b$. Then $(\mathrm{Pas}_{b+1}Q)_{ij} = 0$ for $j > i$, and is equal to $\sum_{\ell=j}^{i}\binom{b-\ell}{i-\ell}(-1)^{\ell-j}\binom{b-j}{\ell-j}$ otherwise. The latter expression evaluates to $\binom{b-j}{i-j}\sum_{\ell=j}^{i}(-1)^{\ell-j}\binom{i-j}{\ell-j} = \binom{b-j}{i-j}(1-1)^{i-j}$, which is 0 if $i > j$, and 1 if $i = j$. □

PROOF OF THEOREM 6.1. For part (i), take $\rho = 2$ and $b = 2k-1$. Consider the two k-mixture sources $\boldsymbol{y} = (y,\alpha)$

and $\boldsymbol{z} = (z, \beta)$ given by Theorem 6.2, which have separation $\frac{1}{2k-1}$ and transportation distance $\frac{1}{2(2k-1)}$. For any $b' \leq 2k-2$, the distributions $D^{\boldsymbol{y},b'}$ and $D^{\boldsymbol{z},b'}$ are *identical* and hence indistinguishable even with infinitely many samples, but the stated reconstruction task would allow us to do precisely this.

For part (ii), set $\rho = 3^{c_E+c_A}$, $b = c_A(2k-1)$, and $\tau = \frac{2}{(2k-1)\rho}$. Let $\boldsymbol{y} = (y, \alpha)$ and $\boldsymbol{z} = (z, \beta)$ be as given by Theorem 6.2 (for this b, ρ), which satisfy the required separation property. Suppose that we can perform the stated reconstruction task using N b-snapshots. Then, we can distinguish between $\boldsymbol{y}$ and $\boldsymbol{z}$ with probability at least $1 - \psi$. But this probability is also upper bounded by $\big[1 + d_{\mathrm{TV}}\big((D^{\boldsymbol{y},b})^N, (D^{\boldsymbol{z},b})^N\big)\big]/2$, where $(D^{\boldsymbol{y},b})^N$ and $(D^{\boldsymbol{z},b})^N$ are the N-fold products of $D^{\boldsymbol{y},b}$ and $D^{\boldsymbol{z},b}$ respectively. Thus, $d_{\mathrm{TV}}\big((D^{\boldsymbol{y},b})^N, (D^{\boldsymbol{z},b})^N\big) \geq 1 - 2\psi$. By Proposition 11 and Lemma 12 in [7]

$$\begin{aligned} N &\geq \frac{1}{4d_{\mathrm{TV}}(D^{\boldsymbol{y},b}, D^{\boldsymbol{z},b})} \ln\left(\frac{1}{1-(1-2\psi)^2}\right) \\ &\geq \frac{\rho^{2k-1}}{8 \cdot 3^b} \ln\left(\frac{1}{4\psi}\right) = \Omega\left(3^{c_E(2k-1)} \ln\left(\frac{1}{\psi}\right)\right) \end{aligned}$$

where the second inequality follows from Theorem 6.2 and Lemma 6.3. □

7. REFERENCES

[1] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proc. 18th COLT*, pages 458–469, 2005.

[2] A. Anandkumar, D.P. Foster, D. Hsu, S.M. Kakade, and Y.-K. Liu. Two SVDs suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation. *CoRR*, abs/1204.6703, 2012. `http://arxiv.org/abs/1204.6703`.

[3] A. Anandkumar, D. Hsu, and S.M. Kakade. A method of moments for mixture models and hidden Markov models. In *Proc. 25th COLT*, pages 33.1–33.34, 2012. `http://arxiv.org/abs/1203.0683`.

[4] S. Arora, R. Ge, and A. Moitra. Learning topic models — going beyond SVD. In *Proc. 53rd FOCS*, pages 2012.

[5] S. Arora and R. Kannan. Learning mixtures of separated nonspherical Gaussians. *Ann. Appl. Prob.*, 15:69–92, 2005.

[6] Y. Azar, A. Fiat, A. R. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proc. 33rd STOC*, pages 619–626, 2001.

[7] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. Sampling algorithms: lower bounds and applications. In *Proc. 33rd STOC*, pages 266–275, 2001.

[8] A. Barvinok. *A Course in Convexity*, volume 54 of *Graduate Studies in Mathematics*. AMS, 2002.

[9] T. Batu, S. Guha, and S. Kannan. Inferring mixtures of Markov chains. In *Proc. 17th COLT*, pages 186–199, 2004.

[10] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *Proc. 51st FOCS*, pages 103–112, 2010.

[11] D.M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[12] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *J. Machine Learning Res.*, 3:993–1022, 2003.

[13] S.C. Brubaker and S. Vempala. Isotropic PCA and affine-invariant clustering. In *Proc. 49th FOCS*, pages 551–560, 2008.

[14] K. Chaudhuri, E. Halperin, S. Rao, and S. Zhou. A rigorous analysis of population stratification with limited data. In *Proc. 18th SODA*, pages 1046–1055, 2007.

[15] K. Chaudhuri and S. Rao. Beyond Gaussians: Spectral methods for learning mixtures of heavy-tailed distributions. In *Proc. 21st COLT*, pages 21–32, 2008.

[16] K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In *Proc. 21st COLT*, pages 9–20, 2008.

[17] S. Chung and K. Murty. Polynomially bounded ellipsoid algorithms for convex quadratic programming. In O. Mangasarian, R. Meyer, and S. Robinson, editors, *Nonlinear Programming*, volume 4, pages 439–485. Academic Press, Orlando, Florida, 1981.

[18] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SICOMP*, 31(2):375–397, 2002.

[19] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler. On learning mixtures of heavy-tailed distributions. In *Proc. 46th FOCS*, pages 491–500, 2005.

[20] S. Dasgupta. Learning mixtures of Gaussians. In *Proc. 40th FOCS*, pages 634–644, 1999.

[21] S. Dasgupta and L.J. Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

[22] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *Proc. 23rd SODA*, pages 1371–1385, 2012.

[23] D. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[24] J. Feldman, R. O'Donnell, and R.A. Servedio. PAC learning mixtures of axis-aligned Gaussians with no separation assumption. In *Proc. 19th COLT*, pages 20–34, 2006.

[25] J. Feldman, R. O'Donnell, and R.A. Servedio. Learning mixtures of product distributions over discrete domains. *SIAM J. Comput.*, 37(5):1536–1564, 2008.

[26] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proc. 12th COLT*, pages 183–192, 1999.

[27] N. Gravin, J. Lasserre, D. Pasechnik, and S. Robins. The inverse moment problem for convex polytopes. *J. Discrete and Computational Geometry*, 48(3):596–621, 2012.

[28] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. 15th UAI*, pages 289–296, 1999.

[29] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *Proc. IJCAI*, pages 688–693, 1999.

[30] R. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[31] A.T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *Proc. 42nd STOC*, pages 553–562, June 2010.

[32] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *SIAM J. Computing*, 38(3):1141–1156, 2008.

[33] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th STOC*, pages 273–282, 1994.

[34] J. Kleinberg and M. Sandler. Using mixture models for collaborative filtering. *J. Comput. Syst. Sci.*, 74:49–69, 2008.

[35] F. McSherry. Spectral partitioning of random graphs. In *Proc. 42nd FOCS*, pages 529–537, October 2001.

[36] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Proc. 51st FOCS*, pages 93–102, 2010.

[37] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. In *Prof. 46th STOC*, pages 366–375, 2005.

[38] V. Y. Pan. Optimal and nearly optimal algorithms for approximating polynomial zeros. *Computers & Mathematics with Applications*, 31(12):97–138, 1996.

[39] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. *J. Comput. Syst. Sci.*, 61(2):217–235, 2000.

[40] Y. Rabani, L.J. Schulman, and C. Swamy. Inference from sparse sampling. `http://www.cs.technion.ac.il/~rabani/Papers/RabaniSS-manuscript.pdf`, 2008.

[41] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. *J. Comput. Syst. Sci.*, 68(4):841–860, 2004.

[42] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. AMS, 2003.

[43] V. H. Vu. Spectral norm of random matrices. In *Proc. 37th STOC*, pages 423–430, 2005.

APPENDIX

A. PROBABILITY BACKGROUND

We use the following large-deviation bounds in our analysis.

LEMMA A.1 (Chernoff bound; Theorem 1.1 in [23]). *Let $X_1, \ldots, X_N$ be independent random variables with $X_i \in [0,1]$ for all i, and $\mu = \left(\sum_i \mathrm{E}\, X_i\right)/N$. Then, $\Pr\left[\left|\frac{1}{N}\sum_i X_i - \mu\right| > \epsilon\right] \le 2e^{-2\epsilon^2 N}$.*

LEMMA A.2 (Bernstein's inequality; Theorem 1.2 in [23]). *Let $X_1, \ldots, X_N$ be independent random variables with $|X_i| \le b$, $\mathrm{E}[X_i] = 0$ for all i, and let $\sigma^2 = \sum_i \mathrm{Var}[X_i]$. Then, $\Pr\left[|\ sum_i X_i| > t\right] \le 2\exp\left(-\frac{t^2}{2(\sigma^2 + bt/3)}\right)$.*

B. SAMPLE-SIZE DEPENDENCE OF [2, 3, 4] ON N FOR ℓ_1-RECONSTRUCTION

We view $P = (p^1, \ldots, p^k)$ as an $n \times k$ matrix. Recall that $r = \sum_{t=1}^k w_t p_t p_t^\dagger$, $A = \sum_{t=1}^k w_t (p^t - r)(p^t - r)^\dagger$, and $M = rr^\dagger + A$. Let $w_{\max} := \max_t w_t$. We consider isotropic k-mixture sources, which is justified by Lemma 3.3. So $\frac{1}{2n} \le r_i \le \frac{2}{n}$ for all $i \in [n]$. Note that $\|r\|_1$, $\|r\|_2^2$, and $\|r\|_\infty$ are all $\Theta\left(\frac{1}{n}\right)$. It will be convenient to split the width parameter ζ into two parameters. Let (i) $\frac{\zeta_1}{\sqrt{n}} = \min_{p,q \in P, p \ne q} \|p - q\|_2$; and (ii) $\zeta_2^2 \|r\|_\infty$ be the smallest non-zero eigenvalue of A. Then, the width of (w, P) is $\zeta = \min\{\zeta_1, \zeta_2\}$. We use $\sigma_i(Z)$ to denote the i-th largest singular value of a matrix Z. If Z has rank ℓ, its condition number is given by $\kappa(Z) := \sigma_1(Z)/\sigma_\ell(Z)$. For a square matrix Z with real eigenvalues, we use $\lambda_i(Z)$ to denote the i-th largest eigenvalue of Z. Note that if Z is an $n \times k$ matrix, then $\sigma_i(Z)^2 = \lambda_i(ZZ^\dagger) = \lambda_i(Z^\dagger Z)$ for all $i = 1, \ldots, k$. Also the singular values of $ZZ^\dagger$ coincide with its eigenvalues, and the same holds for $Z^\dagger Z$.

We now proceed to evaluate the sample-size dependence of [2, 3, 4] on n for reconstructing the mixture constituents within ℓ_1-distance ϵ. Since these papers use different parameters than we do, in order to obtain a meaningful comparison, we relate their bounds to our parameters ζ_1, ζ_2; we keep track of the resulting dependence on n but ignore the (polynomial) dependence on other quantities. We show that the sample size needed is at least $\Omega\left(\frac{n^4}{\epsilon^2}\right)$, with the exception of Algorithm B in [3], which needs $\Omega\left(\frac{n^3}{\epsilon^2}\right)$ samples. As required by [2, 3, 4], we assume that P has full column rank. It follows that M has rank k and A has rank $k-1$. The following inequality will be useful.

PROPOSITION B.1. *Let $D = \mathrm{diag}(d_1, \ldots, d_k)$ where $d_1 \ge d_2 \ge \ldots \ge d_k > 0$. Then $\lambda_k(PDP^\dagger) \ge d_k \lambda_k(PP^\dagger) = d_k \sigma_k(P)^2$.*

Comparison with [4]. The algorithm in [4] requires also that P be ρ-separable. This means that for every $t \in [k]$, there is some $i \in [n]$ such that $p_i^t \ge \rho$ and $p_i^{t'} = 0$ for all $t' \ne t$. This has the following implications. For any $t, t' \in [k]$, $t \ne t'$, we have $\|p^t - p^{t'}\|_2 \ge \sqrt{2}\rho$, so $\frac{\zeta_1}{\sqrt{n}} \ge \sqrt{2}\rho$. We can write $P^\dagger P = Y + Z$, where Y is a PSD matrix, and Z is a diagonal matrix whose diagonal entries are at least ρ^2. So $\lambda_k(P^\dagger P) = \lambda_k(PP^\dagger) \ge \rho^2$. Therefore,

$$\zeta_2^2 \|r\|_\infty + \|r\|_2^2 = \lambda_k(A) + \|r\|_2^2 \ge \lambda_k(M) \ge w_{\min} \cdot \rho^2$$

where the first inequality follows from Lemma 2.1, and the second from Proposition B.1. It follows that $\rho = O\left(\frac{1}{\sqrt{n}}\right)$. The bound in [4] to obtain ℓ_∞ error ε is (ignoring dependence on other quantities) $\Omega\left(\frac{1}{\varepsilon^2 \rho^6}\right)$. So setting $\varepsilon = \frac{\epsilon}{n}$ to guarantee ℓ_1-error at most ϵ and plugging in the above upper bounds on ρ, we obtain that the sample size is $\Omega\left(\frac{n^5}{\epsilon^2}\right)$.

Comparison with [2]. The sample size required by [2] for the latent Dirichlet model for obtaining ℓ_2 error ε is $\Omega\left(\frac{1}{\varepsilon^2 \sigma_k(P)^6}\right)$. Proposition B.1 yields $\lambda_k(M) \ge w_{\min} \cdot \sigma_k(P)^2$ and as argued above, $\lambda_k(M) \le \lambda_k(A) + \|r\|_2^2 = O\left(\frac{1}{n}\right)$. So $\sigma_k(P)^6 = O\left(\frac{1}{n^3}\right)$. Setting $\varepsilon = \frac{\epsilon}{\sqrt{n}}$ for ℓ_1 error ϵ, this yields a bound of $\Omega\left(\frac{n^4}{\epsilon^2}\right)$.

Comparison with [3]. Algorithm A in [3] requires sample size $\Omega\big(\frac{1}{\sigma_k(P)^8\sigma_k(M)^4\varepsilon^2}\big)$ to recover each p^t to within ℓ_2-distance $\varepsilon \max_{p\in P}\|p\|_2$. Since $\max_{p\in P}\|p\|_2 \le \frac{2}{w_{\min}\sqrt{n}}$ due to isotropy, we can set $\varepsilon = \frac{\epsilon w_{\min}}{2}$ to obtain ℓ_1-error ϵ. Since $\sigma_k(P)^2$ and $\sigma_k(M) = \lambda_k(M)$ are both $O\big(\frac{1}{n}\big)$, we obtain a bound of $\Omega\big(\frac{n^8}{\epsilon^2}\big)$.

Algorithm B in [3] uses sample size $\Omega\Big(\kappa(P)^8/\big(\frac{\zeta_1^2}{n}\cdot\sigma_k(M)^2\varepsilon^2\big)\Big)$ to recover each p^t to within ℓ_2-distance $\varepsilon \max_{p\in P}\|p\|_2$. Clearly $\kappa(P) \ge 1$. Again, setting $\varepsilon = \frac{\epsilon w_{\min}}{2}$, this yields a sample size of $\Omega\big(\frac{n^3}{\epsilon^2}\big)$ for ℓ_1 error ϵ.