Learning Arbitrary Statistical Mixtures of Discrete Distributions*

Jian Li[†]
Tsinghua University
Beijing, China 100084.
lapordge@gmail.com

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@uwaterloo.ca

ABSTRACT

We study the problem of learning from unlabeled samples very general statistical mixture models on large finite sets. Specifically, the model to be learned, ϑ, is a probability distribution over probability distributions p, where each such p is a probability distribution over $[n] = \{1, 2, \ldots, n\}$. When we sample from ϑ, we do not observe p directly, but only indirectly and in very noisy fashion, by sampling from $[n]$ repeatedly, independently K times from the distribution p. The problem is to infer ϑ to high accuracy in transportation (earthmover) distance.

We give the first efficient algorithms for learning this mixture model without making any restricting assumptions on the structure of the distribution ϑ. We bound the quality of the solution as a function of the size of the samples K and the number of samples used. Our model and results have applications to a variety of unsupervised learning scenarios, including learning topic models and collaborative filtering.

*A full version is available from the CS arXiv.

[†]Supported in part by the National Basic Research Program of China grants 2015CB358700, 2011CBA00300, 2011CBA00301, and the National NSFC grants 61202009, 61033001, 61361136003. Work performed in part at the Simons Institute for the Theory of Computing.

[‡]Supported by BSF grant number 2012333, and by the Israeli Center of Excellence on Algorithms.

[§]Supported in part by NSF grant 1319745. Work performed in part at the Simons Institute for the Theory of Computing.

[¶]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 ACM 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.
STOC'15, June 14–17, 2015, Portland, OR, USA.
Copyright © 2015 ACM 978-1-4503-3536-2/15/06 ...$15.00.
http://dx.doi.org/10.1145/2746539.2746584.

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; Transportation distance; Kantorovich-Rubinstein duality; Approximation theory; Convex geometry; Spectral methods

1. INTRODUCTION

We study the problem of learning from unlabeled samples a statistical mixture model that is a combination of distributions over a common large discrete domain $[n] = \{1, 2, \ldots, n\}$. This is a model that has applications to a variety of unsupervised learning scenarios, including learning *topic models* [26, 34] and *collaborative filtering* [27]. For instance, in the setting of topic models, we are given a corpus of documents, where each document is a "bag of words" (that is, each document is an unordered multiset of words). The words in a document reflect the topics that this document relates to. The assumption is that there is a small number of "pure" topics, where each topic is a distribution over the underlying vocabulary of n words, and that each document is some combination of topics. Specifically, a K-word document is generated by selecting a "mixed" topic from a probability distribution over convex combinations of pure topics, and then sampling K words from this mixed topic. A good example is the so-called latent Dirichlet allocation model of [10], where the distribution over topic-combinations is the Dirichlet distribution.

The mixture model. In this paper, we consider *arbitrary* such mixtures (of a more general form), and our goal is to learn the mixture distribution, which could be discrete, i.e., have finite support, or continuous. More precisely, the mixture distribution, ϑ, is a probability distribution over probability distributions on $[n]$. (Equivalently, ϑ is a distribution over the $(n-1)$-simplex $\Delta_n = \{x \in \mathbb{R}^n_+ \mid \|x\|_1 = 1\}$.) When we draw a sample from ϑ, we obtain a distribution $p \in \Delta_n$. However, we do not observe p directly, but only indirectly and in very noisy fashion, by sampling K times independently from p. Thus, our sample is a string of length K over

the alphabet $[n]$ where each letter is an iid sample from p. We call such a sample a *K-snapshot* of p. (A k-snapshot corresponds to a document of length K in the topic-model example.) The problem is to learn ϑ with high accuracy.

Our mixture model is more general than that in the topic-model learning example, in that we do not assume that ϑ is supported on the convex hull of k distributions. It is an example of a *statistical mixture model*, where the probability distribution from which the learning algorithm gets samples (the mixed topic generating a document, in our topic-model example) is a mixture of other probability distributions (pure topics, in our example) that are called the mixture *constituents*.

Our results. We give the *first* efficient algorithms for learning a mixture model *without placing any restrictions* on the mixture. We bound the quality of the solution as a function of the size of the samples; clearly, larger samples give better results. A natural way to measure the accuracy of an estimate $\widetilde{\vartheta}$ in our general mixture model is to consider the *transportation distance* (aka *earthmover distance*) between $\widetilde{\vartheta}$ and ϑ (see Section 2) where the underlying metric on distributions over $[n]$ is the L_1 *(or total variation) distance*.

Given a mixture ϑ supported on a k-dimensional subspace, our algorithms return an estimate $\widetilde{\vartheta}$ that is ϵ-close to ϑ in transportation distance, for any $\epsilon > 0$, using K-snapshot samples for $K = K(\epsilon, k)$ and sample size that is $\mathrm{poly}(n)$ and a suitable function of k and ϵ. (Note that the intersection of a k-dimensional subspace with Δ_n could have $\exp(k)$ extreme points; so saying that ϑ lies in a k-dimensional subspace is substantially weaker than assuming that ϑ is supported on the convex hull of k points.) Our main result (Theorem 5.3) is an efficient learning algorithm that uses $O(k^4 n^3 \log n/\epsilon^6)$ 1- and 2- snapshot samples, and $(k/\epsilon)^{O(k)}$ K-snapshot samples, where $K = \widetilde{\Omega}(k^{11}/\epsilon^{10}) = \mathrm{poly}(k, 1/\epsilon)$. We also devise algorithms with different tradeoffs between the sample size and the *aperture*, which is the maximum number of snapshots used per sample point (i.e., document size), for some special cases of the problem. This includes, most notably, the case where ϑ is a *k-spike mixture*, i.e., is supported on k points in Δ_n (Theorem 6.1). This setting has been considered previously (see below), but our algorithm is cleaner and fits into our more general method; and more importantly, our bounds do not depend on distribution-dependent parameters (see the discussion below).

To put our bounds in perspective, first notice importantly that we consider transportation distance with respect to the L_1-metric on distributions. This yields quite strong guarantees on the quality of our reconstruction, however working with the L_1-metric (instead of L_2) makes the reconstruction task much harder, both in terms of technical difficulty (see "Our techniques" below) and the sample-size required: the L_1 distance between two distributions can be much larger than their L_2 distance, so it is much more demanding to bound the L_1-error. In particular, this implies that the sample size must depend on n: as noted in [36], with 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. The sample size needs to depend exponentially on the dimension k because one can have an $\exp(k)$-spike mixture ϑ (on Δ_n) lying in a k-dimensional subspace whose constituents are $\Omega(1)$ L_1-distance apart; recovering an ϵ-close estimate now entails that we isolate the locations of the spikes reasonably accurately, which necessitates $\exp(k)$ sample size. Finally, the aperture must depend on k and ϵ. The dependence on k is simply because our learning task is at least as hard as learning k-spike mixtures for which aperture $2k-1$ is necessary [36]. The dependence on ϵ is because the lower bounds in [36] show that there are two (even single-dimensional) ℓ-spike mixtures, where $\ell = \Theta(1/\epsilon)$, with transportation distance $\Omega(\epsilon)$ that yield identical K-snapshot distributions for all $K < 2\ell - 1$.

A noteworthy feature of *all* our results is that our bounds depend *only* on n, k, and ϵ. In contrast, all previous results for learning topic models (including those that consider only k-spike mixtures) obtain bounds that depend on distribution-dependent parameters such as some measure of the separation between mixture constituents [34, 36], the minimum weight placed on a mixture constituent, and/or the eigenvalues (or singular values) of the covariance matrix (e.g., bounds on σ_k, or L_1-condition numbers, or the robustly simplicial condition) [31, 6, 4, 5]. The distribution-free nature of our bounds is clearly a desirable feature; if the desired accuracy is cruder than the distribution-dependent parameters, then fewer samples are needed.

Our techniques. The main result (Theorem 5.3) is derived as follows. First, we use spectral methods to compute from 1- and 2-snapshot samples a basis B for a subspace $\mathrm{Span}(B)$ of dimension at most k that nearly contains the support of ϑ, and such that learning the projection ϑ_B of ϑ on $\mathrm{Span}(B)$ suffices to learn ϑ (Section 4). We need to choose B carefully so as to overcome various technical challenges that arise because we work with transportation distance in the L_1-metric. Specifically, we need to move between the L_1 and L_2 metrics at various points (the rotational invariance of the L_2-metric makes it easier to work with L_2), and to avoid a $\sqrt{n}$-factor distortion due to this movement, we need to establish that an L_1-ball in $\mathrm{Span}(B)$ is close to being an L_2-ball in $\mathrm{Span}(B)$ (see Lemma 4.5). This allows one to argue that: (a) ϑ_B is supported in an L_2-ball of radius $O\big(\frac{1}{\sqrt{n}}\big)$, which makes it feasible to learn it within L_2-error $\frac{\epsilon}{\sqrt{n}}$ (and hence L_1-error ϵ); and (b) projecting this reconstructed mixture to Δ_n preserves the L_1-error (up to a $\mathrm{poly}\big(k, \frac{1}{\epsilon}\big)$ factor). We remark that the standard SVD technique does not suffice for our purpose, since the resulting subspace need not satisfy the above "spherical" property of L_1-balls (see also the discussion in Section 4). Next, we define a projection of the K-snapshot samples using B. We compute the estimate $\widetilde{\vartheta}_B$ of ϑ_B by averaging the projections and transforming the result to $\mathrm{Span}(B)$ (see Section 5). The proof relies on large deviation bounds. One can show that ϑ is close to ϑ_B. The output $\widetilde{\vartheta}_B$ converges to this projection as the number of samples grows. The rate of convergence can be bounded using tools from approximation theory.

The result for the special case of k-spike mixtures (i.e., ϑ is supported on k distributions) uses a three-step approach analogous to the argument in [36], but the implementation of each step is different). The first step finds B as in the general case. In the second step, the algorithm projects the sample data onto the basis vectors in B. From this data, the algorithm computes a good approximation to the projection of ϑ onto each axis. The idea is to use linear programming to compute a piecewise constant discretization of the projected measure such that the first K moments are close to the empirical moments derived from the samples of K-snapshots.

The analysis uses a classical result in approximation theory due to Jackson that estimates the error in approximating a 1-Lipschitz function on $[0,1]$ by the first K Chebyshev polynomials. (In fact, this step, too, does not use the special structure of the mixture. It works in the case of an arbitrary measure ϑ, and our error estimates are asymptotically optimal in general.) In the third step, we use the approximate projected measures to compute a good approximation for the projection of ϑ on $\mathrm{Span}(B)$, giving our algorithm's output. The main idea here is similar to that of the second step. We discretize the projection and use a linear program to compute a discretized measure whose projections onto the axes used in the second step give a good match to the computed approximations on those axes. The analysis of this algorithm uses Yudin's multidimensional generalization of Jackson's theorem [42]. Both the second step and the third step use Kantorovich-Rubinstein duality to relate the results from approximation theory to the approximation guarantees in terms of the transportation distance.

Related work. Generally speaking, our problem is an example of learning a *mixture model*. Unlike our case, other mixture learning problems, such as learning a mixture of Gaussians (see [18, 9, 32]), assume a special structure of the distributions that contribute to the mixture. We discuss this related literature below.

A few previous papers consider the problem of learning a topic model [6, 3, 4, 36]. They all make limiting assumptions on the structure of the mixture model. The only paper that considers an arbitrary distribution ϑ over combinations of topics is [6]. However, this paper assumes that the pure topics are ρ-separated, which means that each topic has an anchor word that has probability at least ρ in this topic, and probability 0 in any other topic. In the case of an arbitrary ϑ (over such topics), the paper [6] learns the correlation matrix for pairs of pure topics and not ϑ. In the special case of latent Dirichlet allocation, the paper also reconstructs ϑ. The latent Dirichlet allocation setting is also considered in [3]. For this special case, they relax the condition in [6] to the requirement that the matrix whose columns are the word distributions of the k pure topics has full rank k. The constraints on the model that are imposed in [6, 3] allow them to achieve their learning goals using documents of constant size that is independent of the number of pure topics k and the desired accuracy ϵ. As we show in this paper, this is impossible in the general case. The remaining two papers mentioned above [4, 36] consider only the case where each document is generated from a single pure topic, so ϑ is a discrete distribution with support of size k. The first paper [4] imposes on the pure topics the same rank condition as in [3], and thus is able to learn the model from constant size documents. The second paper [36] studies the general pure topic documents case and shows how to learn the model from documents of size $2k-1$, which is a tight requirement. Notice that in this case, the document size is independent of the desired accuracy. Our results specialized to this case are motivated by the techniques in [36]. They give a simpler and cleaner proof that roughly matches the results there (in particular, the mixture model is recovered using K-snapshots for $K = 2k-1$, which is optimal).

Learning statistical mixture models has been studied in the theory community for about twenty years. The defining problem of this area was the problem of learning a mixture of high-dimensional Gaussians. Starting with the groundbreaking result of [18], a sequence of improved results [19, 7, 40, 29, 1, 23, 12, 28, 9, 32] resolved the problem. Beyond Gaussians, various recent papers analyze learning other highly structured mixture models (e.g., mixtures of discrete product distributions) [30, 25, 16, 8, 33, 17, 24, 29, 13, 15, 14, 20]. An important difference between this work and ours is that the structure of the mixtures that they discuss enables learning using samples that consist of a 1-snapshot of a random mixture constituent (which is impossible in our setting). Since Gaussians and other structured mixtures can be learned from 1-snapshot samples, the issue of the samples themselves being generated from a combination of the mixture constituents does not arise there. Our problem is unique to learning from multi-snapshot samples.

2. PRELIMINARIES AND NOTATION

Let $T : X \to Y$ be a transformation from a normed space X (with norm $\|\cdot\|_X$) to a normed space Y (with norm $\|\cdot\|_Y$). Let μ be a measure defined over X. We use $\mu \circ T^{-1}$ to denote the image measure (or pushforward measure) defined over Y: $\mu \circ T^{-1}(U) = \mu(T^{-1}(U))$ for all measurable $U \subset Y$. It is a simple fact that (see e.g., [22]) that for any measurable function f,

$$\int_Y f \,\mathrm{d}(\mu \circ T^{-1}) = \int_X f \circ T \,\mathrm{d}\mu. \tag{1}$$

For ease of notation, we sometimes write $T\mu$ to denote the image measure $\mu \circ T^{-1}$. For a vector v, we use $\|v\|$ to denote its L_2 norm, and for an operator T, we use $\|T\|_{X\to Y}$ to denote its operator norm (i.e., $\|T\|_{X\to Y} = \sup\{\|Tx\|_Y \mid x \in X, \|x\|_X = 1\}$). For ease of notation, we use $\|T\|$ to denote the $L_2 \to L_2$ operator norm of T.

Transportation Distance. Let (X, d) be a separable metric space. Recall that for any two distributions P and Q on S, the transportation distance $\mathrm{Tran}(P, Q)$ (also called Rubinstein distance, Wasserstein distance or earth mover distance in literature) is defined as

$$\mathrm{Tran}(P,Q) := \inf\left\{\int d(x,y)\,\mathrm{d}\mu(x,y) : \mu \in M(P,Q)\right\} \tag{2}$$

where $M(P, Q)$ is the set of all joint distributions (also called coupling) on $X \times X$ with marginals P and Q. For the discrete case (say X is a finite set of discrete points $v_1, \ldots, v_n$), (2) is in fact the following familiar transportation LP:

$$\min \sum_{i,j} d(v_i, v_j) x_{ij} \quad \text{s.t.} \quad \sum_j x_{ij} = P(\{v_i\})\ \forall i \in [n],$$
$$\sum_i x_{ij} = Q(\{v_j\})\ \forall i \in [n], \qquad x_{ij} \in [0,1]\ \forall i,j \in [n].$$

Any feasible solution $\{x_{ij}\}_{i,j}$ to the above LP is in fact a coupling of P and Q, since it can be interpreted as a joint distribution over $X \times X$, and the constraints of the LP dictate the first marginal of $\{x_{ij}\}$ is P and the second is Q.

Suppose μ is a measure on some metric space (X, d). Let $T : X \to X$ be an operator. T naturally defines a coupling W between μ and the image measure $T\mu$: for any $R \subseteq X \times X$, let $W(R) = \mu(\{x \mid (x, Tx) \in R\})$ (so for any measurable $S \subseteq X$, $W(S \times T(S)) = \mu(S)$). For ease of description, for such a coupling, we often say "we couple x with Tx together".

Let 1-Lip be the set of 1-Lipschitz functions on X, i.e., 1-$\mathsf{Lip} := \{f : X \to \mathbb{R} \mid |f(x) - f(y)| \le d(x,y)$ for any $x, y \in$

$X\}$. We need the following important theorem by Kantorovich and Rubinstein (see e.g., [22]):

$$\mathrm{Tran}(P,Q) = \sup\left\{\left|\int f \mathrm{d}(P-Q)\right| : f \in 1\text{-}\mathsf{Lip}\right\}. \quad (3)$$

In the discrete case, Kantorovich-Rubinstein theorem is exactly LP-duality: the dual of the aforementioned LP is

$$\max \sum_i f_i(P(\{v_i\}) - Q(\{v_i\})) \text{ s.t. } f_i - f_j \le d(v_i, v_j)\ \forall i,j \in [n].$$

It is important to notice the transportation distance and the Lipschitz condition are associated with the same metric $d(x,y)$. We use Tran_1 and Tran_2 to denote the transportation distance for L_1 and L_2 metrics respectively. In 1-dimensional space, L_1 and L_2 are the same and we simply use Tran. The following simple lemma will be useful in several places. The proofs are standard and are deferred to the full version.

LEMMA 2.1. *$(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are two normed spaces. We are given two probability measures μ, ν defined over X such that $\mathrm{Tran}(\mu,\nu) \le \epsilon$.*

(i) *Suppose $T : X \to Y$ is a transformation from X to Y. $\mathrm{Tran}(T\mu, T\nu) \le \epsilon \cdot \|T\|_{X\to Y}$.*

(ii) *Furthermore, if both μ and ν are supported on a subspace $V \subset X$, then $\mathrm{Tran}(T\mu, T\nu) \le \epsilon \cdot \|T\|_V$, where $\|T\|_V = \sup_{x\in V} \|Tx\|_Y / \|x\|_X$.*

(iii) *We are given two operators T and T' such that $\|T - T'\|_{X\to Y} \le \epsilon$. Suppose $\|T\|_{X\to Y} = O(1)$ and $\|x'\|_X = O(1)$ for all $x' \in \mathrm{Support}(\nu)$. Then, we have that $\mathrm{Tran}(T\mu, T'\nu) \le O(\epsilon)$.*

We state the following standard Chernoff-Hoeffding bound and Bernstein inequality.

PROPOSITION 2.2. *Let $\{X_i\}_{i\in[n]}$ be independent random variables. Let $X = \sum_{i=1}^n X_i$, and $t > 0$ be arbitrary.*

(i) *Suppose $X_i \in [0,1]$ for all i. Then, we have that $\Pr\left[\,|X - \mathbb{E}[X]| > t\right] < 2\exp(-2t^2/n)$.*

(ii) *Suppose $|X_i| \le 1$, $\mathbb{E}[X_i] = 0$ for all i. Let $\sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[X_i]$. Then, $\Pr\left[\,|X| > t\right] \le 2\exp\left(-\frac{t^2}{2(\sigma^2 + t/3)}\right)$.*

We will use the following results from the matrix perturbation and random matrix theory.

THEOREM 2.3. *(Wedin's theorem, see e.g., [38, pp.261]) Let $A, \widetilde{A} \in \mathbb{R}^{m\times n}$ with $m \ge n$ be given. Let the singular value decompositions of A and $\widetilde{A}$ be*

$$(U_1, U_2, U_3)^T A (V_1, V_2) = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{pmatrix},$$

$$(\widetilde{U}_1, \widetilde{U}_2, \widetilde{U}_3)^T \widetilde{A} (\widetilde{V}_1, \widetilde{V}_2) = \begin{pmatrix} \widetilde{\Sigma}_1 & 0 \\ 0 & \widetilde{\Sigma}_2 \\ 0 & 0 \end{pmatrix}$$

Let Φ be the matrix of canonical angles between $\mathrm{Span}(U_1)$ and $\mathrm{Span}(\widetilde{U}_1)$ and Θ be that between $\mathrm{Span}(V_1)$ and $\mathrm{Span}(\widetilde{V}_1)$. If there exists $\delta, \alpha > 0$ such that $\min_i \sigma_i(\widetilde{\Sigma}_1) \ge \alpha + \delta$ and $\max_i \sigma_i(\Sigma_2) \le \alpha$, then $\max\{\|\sin\Phi\|, \|\sin\Theta\|\} \le \frac{\|A - \widetilde{A}\|}{\delta}$. Moreover, $\|\Pi_A - \Pi_{\widetilde{A}}\| = \|\sin\Phi\|$ (see e.g., [38, pp.43]).

THEOREM 2.4 ([41]). *For every constant $c > 0$, there is a constant $C > 0$ such that the following holds. Let A be a symmetric with entries $a_{ij} = a_{ji} = X_{ij}$, where X_{ij}, $1 \le i \le j \le n$ are independent random variables. Suppose each X_{ij} is such that $|X_{ij}| < K$, $\mathbb{E}[X_{ij}] = 0$ and $\mathrm{Var}[X_{ij}] \le \sigma^2$ where $\sigma \ge C^2 K \ln^2 n/\sqrt{n}$. Then, it holds that*

$$\Pr[\|A\| \le 2\sigma\sqrt{n} + C(K\sigma)^{1/2} n^{1/4} \ln n] \ge 1 - 1/n^c.$$

The Chebyshev polynomial (of the first kind) is defined as the polynomial satisfying $T_n(\cos(x)) = \cos(nx)$. An equivalent recursive definition is: $T_0(x) = 1, T_1(x) = x$ and $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$. We need the classical Jackson's theorem (see e.g., [37]) in approximation theory (specialized to our setting) and a multidimensional generalization of Jackson's theorem established by Yudin [42] (Theorem 2.6).

THEOREM 2.5 (JACKSON'S THEOREM). *It is possible to approximate any function on $[0,1]$ in $1\text{-}\mathsf{Lip}$ within L_∞ error $O(1/K)$ using Chebyshev polynomials (or equivalently trigonometric polynomials) of degree at most K, i.e., there exist $\{t_i\}_{i\in[K]}$ such that $f(x) = \sum_{i=0}^K t_i T_i(x) \pm O(1/K) \quad \forall x \in [0,1]$. Moreover, $|t_i| \le \mathrm{poly}(K)$ for all $i = 0, \ldots, K$.*

THEOREM 2.6. *We use $\mathsf{B}_2^h(R)$ to denote the sphere $\{x \in \mathbb{R}^h \mid \|x\|_2 \le R\}$. For any function $f : \mathsf{B}_2^h(1) \to \mathbb{C}$ which is $1\text{-}\mathsf{Lip}$ (in L_2 distance), there exists complex numbers $c(t')$ for $t' \in \mathbb{Z}^h \cap \mathsf{B}_2^h(R)$, such that $|c(t')| \le \exp(O(h))$[1] and for all $x \in \mathsf{B}_2^h(1)$,*

$$\left| f(x) - \sum_{t' \in \mathbb{Z}^h \cap \mathsf{B}_2^h(R)} c(t') e^{\mathrm{i}\langle t', x\rangle} \right| \le O\left(\frac{h}{R}\right).$$

3. LEARNING SINGLE-DIMENSIONAL MIXTURES: THE COIN PROBLEM

In this section, we consider the problem of learning a mixture ϑ supported on $[0,1]$, which we call the *coin problem*. Using results in [36], these results carry over to the setting where ϑ supported on a line segment in the $(n-1)$-simplex $\Delta_n = \{x \in \mathbb{R}^n_{\ge 0}, \|x\|_1 = 1\}$. We first consider an arbitrary (even continuous) ϑ in $[0,1]$; in Section 3.1, we consider the case where ϑ is a *k-spike mixture*.

Let $B_{i,K}(x) = \binom{K}{i} x^i (1-x)^{K-i}$. Let N_K denote the number of K-snapshots we take from ϑ. For $0 \le i \le K$, define $\mathsf{fq}_i(\vartheta) := \int B_{i,K}(x)\mathrm{d}\vartheta$. We call $\mathsf{fq}(\vartheta) := \{\mathsf{fq}_i(\vartheta)\}_{0\le i\le K}$ the *frequency vector* corresponding to ϑ. We use $\widetilde{\mathsf{fq}}_i$ to denote the fraction of sampled coins that showed "heads" exactly i times and let $\widetilde{\mathsf{fq}} := \{\widetilde{\mathsf{fq}}_i\}_{0\le i\le K}$ be the *empirical frequency vector*. It is easy to see that $\mathsf{fq}(\vartheta) = \mathbb{E}[\widetilde{\mathsf{fq}}]$. If we take enough samples, the frequency vector corresponding to the empirical measure $\widetilde{\vartheta}$ should be sufficiently close to that of ϑ.

LEMMA 3.1. *By taking $N_K = \kappa^{-2}\log(K/\delta)$ samples, with high probability $1-\delta$, we have that $\|\,\mathsf{fq}(\vartheta) - \widetilde{\mathsf{fq}}\,\|_\infty \le \kappa$.*

THEOREM 3.2. *There exists an algorithm, with running time polynomial in K, that gets as input $m = \mathrm{poly}(K)$ coins*

[1] In Yudin's theorem, $c(t')$ is in fact $\hat{f}(t')\lambda(t'/R)$, where $\hat{f}(t') = \frac{1}{(2\pi)^h}\int_{x\in[-\pi,\pi]^d} f(x) e^{-\mathrm{i}\langle t', x\rangle}\,\mathrm{d}x$ is the Fourier coefficient, $\lambda(x) = (\phi * \phi)(x)$, $\phi(x)$ is the first normalized eigenfunction of a PDE known as Helmholtz equation, and $*$ is the convolution.

from a mixture ϑ, each tossed K times, and output a mixture $\widehat{\vartheta}$ such that $\mathrm{Tran}(\vartheta, \widehat{\vartheta}) \le O(1/\sqrt{K})$ *with high probability.*

Theorem 3.2 can be proved by a simple application of Chernoff bound (where we set $\widehat{\vartheta}(\{\frac{i}{K}\}) = \widetilde{\mathsf{fq}}_i$), which we omit here. We provide an alternative proof based on Bernstein polynomials later. It is a natural question to ask whether $O(1/\sqrt{K})$ in Theorem 3.2 achieves the optimal aperture-transportation distance tradeoff. In [36], it is shown that recovering a K-spike mixture within transportation distance $O(1/K)$ using $c(2K-1)$ (for any constant $c \ge 1$) aperture requires $\exp(\Omega(K))$ samples. The following theorem provides a *matching upper bound.*

THEOREM 3.3. *There exists an algorithm, with running time polynomial in K, that gets as input $m = \exp(O(K))$ coins from a mixture ϑ, each tossed K times, and outputs a mixture $\widehat{\vartheta}$ such that* $\mathrm{Tran}(\vartheta, \widehat{\vartheta}) \le O(1/K)$ *with high probability.*

To prove Theorem 3.3, we make a crucial observation (Lemma 3.4) that links the transportation distance, the frequency vector and the coefficients of Bernstein polynomial approximation. Lemma 3.6 bounds these coefficients using the relation between Bernstein polynomial basis and Chebyshev polynomial basis. We then provide a simple LP-based algorithm to reconstruct ϑ.

LEMMA 3.4. *Suppose for any $f \in 1\text{-}\mathsf{Lip}[0,1]$, there exist $K+1$ real numbers $c_0, \ldots, c_K \in [-C, C]$, for some value $C > 0$ and $\lambda > 0$, such that $f = \sum_i c_i B_{i,K} \pm O(\lambda)$. Then for any two distributions P and Q on $[0,1]$,* $\mathrm{Tran}(P, Q) \le C \cdot \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 + O(\lambda)$.

PROOF. We have $\mathsf{fq}_i(P) = \int B_{i,K} \, \mathrm{d}P$. For any $f \in 1\text{-}\mathsf{Lip}$ such that $f(x) \in [0,1]$ for all $x \in [0,1]$, we have

$$\begin{aligned}\left|\int f \mathrm{d}(P-Q)\right| &= \left|\sum_{i=0}^{K} c_i \int B_{i,K} \, \mathrm{d}(P-Q)\right| + O(\lambda) \\ &= \left|\sum_{i=0}^{K} c_i (\mathsf{fq}_i(P) - \mathsf{fq}_i(Q))\right| + O(\lambda) \\ &\le C \cdot \|\mathsf{fq}(P) - \mathsf{fq}(Q)\|_1 + O(\lambda).\end{aligned}$$

Taking supremum over f on both sides of the above inequality yields the lemma. ☐

LEMMA 3.5. *For any function $f \in 1\text{-}\mathsf{Lip}[0,1]$, there exists $K+1$ real numbers $c_0, \ldots, c_K \in [-C, C]$ with $C = O(1)$ such that $f(x) = \sum_{i=0}^{K} c_i B_{i,K}(x) \pm O(1/\sqrt{K})$ for all $x \in [0,1]$.*

LEMMA 3.6. *For any function $f \in 1\text{-}\mathsf{Lip}[0,1]$, there exists $K+1$ real numbers $c_0, \ldots, c_K \in [-C, C]$ with $C = \mathrm{poly}(K) \cdot 2^K$ such that $f(x) = \sum_{i=0}^{K} c_i B_{i,K}(x) \pm O(1/K)$ for all $x \in [0,1]$.*

PROOF. By Jackson's theorem (Theorem 2.5) in approximation theory, for any function $f \in 1\text{-}\mathsf{Lip}[0,1]$, there exist t_is (with $|t_i| \le \mathrm{poly}(K)$) such that $f(x) = \sum_{i=0}^{K} t_i T_i(x) \pm O(1/K)$ for all $x \in [0,1]$, where T_is are Chebyshev polynomials of degrees at most K. Let M be the linear transformation from the $\{T_i\}_{i \in [K]}$ basis to the $\{B_{i,K}\}_{i \in [K]}$ basis. For an arbitrary polynomial $P(x)$ of degree at most K, we can write $P(x) = \sum_{i=0}^{K} c_i B_{i,K}(x) = \sum_{i=0}^{K} t_i T_i(x)$, where $c_i = \sum_{k=0}^{K} M_{ik} t_k$. Using $t = (t_0, \ldots, t_K)^T$ and $c = (c_0, \ldots, c_K)^T$, we have that $c = Mt$. It is known that $|M_{ij}| \le |M_{iK}|$ for all i, j and $|M_{iK}| = (2K-1)!!/(2i-1)!!(2K-2i-1)!!$ where $n!! = n(n-2)(n-4)\ldots(4)(2)$ for even n and $n!! = n(n-2)(n-4)\ldots(3)(1)$ for odd n [35]. Hence, we have that

$$\begin{aligned}\|c\|_\infty \le \|M\|_{\infty \to \infty} \|t\|_\infty &= \left(\max_{0 \le j \le K} \sum_{i=0}^{K} |M_{ij}|\right) \|t\|_\infty \\ &\le K \cdot \frac{(2K-1)!!}{(2i-1)!!(2K-2i-1)!!} \le \mathrm{poly}(K) \cdot 2^K.\end{aligned}$$

This implies that for any $f \in 1\text{-}\mathsf{Lip}$, we can also get c_is with $|c_i| \le \mathrm{poly}(K) 2^K$ such that $f(x) = \sum_{i=0}^{K} t_i T_i(x) \pm O(1/K) = \sum_{i=0}^{K} c_i B_{i,K}(x) \pm O(1/K)$ for all $x \in [0,1]$. ☐

Reconstructing ϑ. Suppose we have a good empirical frequency vector $\widetilde{\mathsf{fq}}$ which satisfies $\|\widetilde{\mathsf{fq}} - \mathsf{fq}(\vartheta)\|_1 \le \lambda/C$, where λ and C are as in Lemma 3.4 Now, we show how to reconstruct the mixture ϑ approximately. We propose a simple LP-based algorithm as follows.

We approximate each $B_{i,K}$ by a piecewise constant function $\overline{B}_{i,K}$ in $[0,1]$ such that $\|B_{i,K} - \overline{B}_{i,K}\|_\infty \le \epsilon'$ for $\epsilon' = O(\kappa)$ (κ in Lemma 3.1). It is easy to see that $O(1/\epsilon')$ pieces suffice (since $B_{i,K}$ is either monotone or unimodal). We can divide $[0,1]$ into $h = O(K/\epsilon')$ small intervals $[a_0 = 0, a_1), [a_1, a_2), \ldots, [a_{h-1}, a_h = 1]$ such that in each small interval $\overline{B}_{i,K}$ is a constant for all $0 \le i \le K$. We use $b_{i,j}$ to denote the value of $\overline{B}_{i,K}$ in interval $[a_j, a_{j+1})$. For each small interval $[a_j, a_{j+1})$, define an variable z_j (think of z_j as the approximation of $\vartheta([a_j, a_{j+1}))$). Consider the following linear program LP:

$$z \ge 0, \quad \sum_{j=0}^{h-1} z_j = 1, \quad \sum_{j=0}^{h-1} b_{i,j} z_j = \widetilde{\mathsf{fq}}_i \pm \epsilon' \ \text{ for all } i = 0, \ldots, K.$$

It is easy to see that, by Lemma 3.1, $z_j = \vartheta([a_j, a_{j+1}))$ defined by the original mixture measure ϑ is a feasible solution for LP.

On the other hand, any feasible solution of LP produces a frequency vector that is close to $\widetilde{\mathsf{fq}}$: Suppose $z^\star$ is an arbitrary feasible solution of LP and $\widehat{\vartheta}$ is any distribution supported on $[0,1]$ that is consistent with $z^\star$ (i.e., $\widehat{\vartheta}([a_j, a_{j+1})) = z_j^\star$), we have that

$$\begin{aligned}\mathsf{fq}_i(\widehat{\vartheta}) &= \int B_{i,K} \mathrm{d}\widehat{\vartheta} = \pm\epsilon' + \int \overline{B}_{i,K} \mathrm{d}\widehat{\vartheta} \\ &= \pm\epsilon' + \sum_j b_{i,j} \int_{[a_j, a_{j+1})} \mathrm{d}\widehat{\vartheta} = \pm\epsilon' + \sum_j b_{i,j} z_i^\star = \widetilde{\mathsf{fq}}_i \pm 2\epsilon'.\end{aligned}$$

PROOF OF THEOREM 3.3. Combining the above bound with Lemma 3.1, we have that

$$\begin{aligned}\|\mathsf{fq}(\widehat{\vartheta}) - \mathsf{fq}(\vartheta)\|_1 &\le K \|\mathsf{fq}(\widehat{\vartheta}) - \mathsf{fq}(\vartheta)\|_\infty \\ &\le K(\|\mathsf{fq}(\widehat{\vartheta}) - \widetilde{\mathsf{fq}}\|_\infty + \|\widetilde{\mathsf{fq}} - \mathsf{fq}(\vartheta)\|_\infty) \le O(K\kappa).\end{aligned}$$

Then, taking $\kappa = O(1/CK^2)$ (recall that $C = \mathrm{poly}(K) 2^K$), using Lemma 3.1 with $2^{O(K)}$ samples, we can make $\| \mathsf{fq}(\widehat{\vartheta}) - \mathsf{fq}(mix) \|_1 \le 1/CK$. So by Lemma 3.6, we finally have that

$$\mathrm{Tran}(\widehat{\vartheta}, \vartheta) \le C \|\mathsf{fq}(\widehat{\vartheta}) - \mathsf{fq}(\vartheta)\|_1 + O(1/K) \le O(1/K) \quad ☐$$

PROOF OF THEOREM 3.2. The proof is the same as that of Theorem 3.3, except that we use Lemma 3.5 instead. In

this case, it suffices to use only poly(K) samples to ensure that $\| \mathsf{fq}(\widehat{\vartheta}) - \mathsf{fq}(\vartheta) \|_1 \le O(1/K)$. $\square$

3.1 Learning k-spike mixtures

We now consider the case where ϑ is a k-spike mixture supported in $[0,1]$, i.e., is supported on k points in $[0,1]$. This result will be useful later when we consider mixtures in higher dimensions. We now use K-snapshots only for $K = 2k-1$. Let the i-th moment of ϑ be $g_i(\vartheta) = \int x^i \vartheta(\mathrm{d}x) = \sum_{j=1}^k p_j \alpha_j^i$. The algorithm is based on an identifiability lemma proved in [36] (Lemma 3.7) and its converse (Lemma 3.8).

LEMMA 3.7 ([36]). *For any two k-spike distributions ϑ_1, ϑ_2 supported on $[0,1]$, $\|g(\vartheta_1) - g(\vartheta_2)\|_2 \ge \left(\frac{\mathrm{Tran}(\vartheta_1,\vartheta_2)}{k}\right)^{O(k)}$.*

LEMMA 3.8. *For any two distributions ϑ_1, ϑ_2 supported on $[0,1]$, and $i \in [K]$, $|g_i(\vartheta_1) - g_i(\vartheta_2)| \le i \cdot \mathrm{Tran}(\vartheta_1, \vartheta_2))$.*

Recall the frequency vector $\mathsf{fq}_i(\vartheta) = \int \binom{K}{i} x^i (1-x)^{K-i} \vartheta(\mathrm{d}x) = \sum_{j=1}^k p_j \binom{K}{i} x^i (1-x)^{K-i}$. Define the *normalized frequency vector* to be $\mathsf{nfq}_i(\vartheta) = \int x^i (1-x)^{K-i} \vartheta(\mathrm{d}x) = \sum_{j=1}^k p_j x^i (1-x)^{K-i}$. Let Pas be the $2k \times 2k$ lower triangular Pascal triangle matrix with non-zero entries $\mathrm{Pas}_{ij} = \binom{K-i}{j-1}$ for $0 \le i \le K$ and $i \le j \le K$. It is not difficult to verify that $g(\vartheta) = \mathrm{Pas}\, \mathsf{nfq}(\vartheta)$. It is known that $\|\mathrm{Pas}\| \le 4^k/\sqrt{3}$. By Lemma 3.1, using $O((k/\epsilon)^{O(k)})$ samples, the empirical frequency vector $\widetilde{\mathsf{fq}}$ satisfies that $\|\widetilde{\mathsf{fq}} - \mathsf{fq}(\vartheta)\|_2 \le (\epsilon/k)^{O(k)}$ with probability 0.99. Let $\widetilde{\mathsf{nfq}}_i = \widetilde{\mathsf{fq}}/\binom{K}{i}$. Let $\widetilde{g} = \mathrm{Pas}\, \widetilde{\mathsf{nfq}}$ be the empirical moment vector.

If we can find a distribution $\widetilde{\vartheta}$ such that $\|g(\widetilde{\vartheta}) - g(\vartheta)\|_2 \le (\epsilon/k)^{\Omega(k)}$, we know, by Lemma 3.7, that $\mathrm{Tran}(\widetilde{\vartheta}, \vartheta) \le \epsilon$. In order to find such a $\widetilde{\vartheta}$, we do the following. $\widetilde{\vartheta}$ is a k-spike distribution supported on the set of discrete points $\{0, \tau, 2\tau, \ldots, 1\}$ where $\tau = (\epsilon/k)^{\Omega(k)}$. First, we guess the support of $\widetilde{\vartheta}$ (there are $\binom{1/\tau}{k}$ choices). Then, we solve the following linear program LP_1, where x_j represents the probability mass placed at point $j\tau \in \mathrm{Support}(\widetilde{\vartheta})$:

$$\mathsf{LP}_1: \quad \left|\sum_j x_j (j\tau)^i - \widetilde{g}_i\right| \le O(K\tau), \text{ for all } i \in [K],$$
$$\sum_j x_j = 1, \qquad x_j \in [0,1], \text{ for all } j$$

THEOREM 3.9. *Using $(k/\epsilon)^{O(k)} \log(1/\delta)$ many $(2k-1)$-snapshot samples, the above algorithm can produce an estimation $\widetilde{\vartheta}$, which satisfies that $\mathrm{Tran}(\widetilde{\vartheta}, \vartheta) \le \epsilon$ with probability $1-\delta$.*

PROOF. We know there is a k-spike measure ϑ' supported on $\{0, \tau, 2\tau, \ldots, 1\}$ such that $\mathrm{Tran}(\vartheta, \vartheta') \le \tau$. Hence, $|g_i(\vartheta') - g_i(\vartheta)| < i\tau$ for all i, by Lemma 3.8. Also,

$$\|\widetilde{g} - g(\vartheta)\|_2 \le \|\mathrm{Pas}\| \|\widetilde{\mathsf{nfq}} - \mathsf{nfq}(\vartheta)\|_2 \le \|\mathrm{Pas}\| \|\widetilde{\mathsf{fq}} - \mathsf{fq}(\vartheta)\|_2$$

which is at most $\left(\frac{\epsilon}{k}\right)^{\Omega(k)}$. Therefore, we have

$$|g_i(\vartheta') - \widetilde{g}_i| \le |g_i(\vartheta') - g_i(\vartheta)| + |g_i(\vartheta) - \widetilde{g}_i| \le O(i\tau).$$

Thus, LP_1 is feasible. Since $\widetilde{\vartheta}$ is a feasible solution of LP_1, we have $\|g(\widetilde{\vartheta}) - \widetilde{g}\|_2 \le O(K^{3/2}\tau)$. So $\|g(\widetilde{\vartheta}) - g(\vartheta)\|_2$ is at most

$$\|g(\widetilde{\vartheta}) - \widetilde{g}\|_2 + \|g(\vartheta) - \widetilde{g}\|_2 \le O(K^{3/2}\tau) \le \left(\frac{\epsilon}{k}\right)^{\Omega(k)}.$$

This implies the theorem, by Lemma 3.7. $\square$

4. LEARNING MULTIDIMENSIONAL MIXTURES ON Δ_N: A REDUCTION

We now consider the setting where the mixture ϑ (on Δ_n) is an arbitrary distribution supported in a k-dimensional subspace in $\mathbb{R}^n$. In this section, we use Tran_1 and Tran_2 to denote the transportation distances measured in L_1 and L_2 norm respectively. For a point v and a set S, we use $\Pi_S(v)$ to denote the projection of v to S, i.e., the point in S that is closest to v. We always assume the projection is with respect to L_2 distance, unless specified otherwise. For any arbitrary measure ϑ supported on $\mathbb{R}^n$, we use $\Pi_S(\vartheta)$ to denote the projected measure defined as $\Pi_S(\vartheta)(T) = \vartheta(\Pi_S^{-1}(T))$ for any measurable $T \subseteq S$.

This section provides a reduction from the original learning problem to to the problem of learning the projected measure in a specific subspace $\mathrm{Span}(B)$. Sections 5 and 6 complement this reduction by devising algorithms for learning the projected measure $\vartheta_B := \Pi_{\mathrm{Span}(B)}(\vartheta)$ (for arbitrary k-dimensional ϑ and k-spike ϑ respectively); combining these algorithms with the reduction of this section yields algorithms for learning ϑ. The space $\mathrm{Span}(B)$ will satisfy several useful properties (Lemma 4.5). One particularly useful property is that any unit vector $v \in \mathrm{Span}(B)$ has $\|v\|_\infty \le O(1/\sqrt{n})$ (ignoring factors depending ϵ and k). This implies that L_1 norm and L_2 norm in $\mathrm{Span}(B)$ are quite close up to scaling, hence allow us to convert bounds between L_1 and L_2 distances without losing a factor depending on n (otherwise, we typically lose a factor of $\sqrt{n}$). Furthermore, we can show we do not lose too much by working in $\mathrm{Span}(B)$ as most of the mass of ϑ is very close to $\mathrm{Span}(B)$. Suppose we can learn the projected measure ϑ_B well. If we can show ϑ_B is close to the original mixture ϑ in Tran_1 distance, then $\widetilde{\vartheta}_B$, a good estimation of ϑ_B, would be a good estimation of ϑ as well. However, we are not able to show ϑ_B and ϑ are close enough in general. Nevertheless, we can prove that a projection of ϑ_B to a smaller polytope is close to ϑ. Finally, we need to make some small adjustments in order to ensure that our estimation $\widetilde{\vartheta}$ is a valid mixture, as well as a good approximation of ϑ (see Reduction 1).

Before we delve into the details of our reduction, we provide some intuition for why we require the subspace $\mathrm{Span}(B)$ to satisfy the above-mentioned properties and why the standard SVD method does not suffice. For ease of discussion, we treat ϵ and k as constants, but n as a parameter that can be very large. Our goal is to obtain $\mathrm{Span}(B)$ of dimension at most k so that if we can learn the projected mixture ϑ_B within Tran_1-distance at most ϵ_1, then we can learn ϑ within Tran_1-distance at most ϵ. We would like ϵ_1 to be independent of n so that the number of K-snapshot samples required to estimate ϑ_B within Tran_1-distance at most ϵ_1 is independent of n (as is the case in Theorems 5.3 and 6.1).

Suppose first that we know A exactly and we simply use $\mathrm{Span}(A)$ as the subspace. In fact, it is not difficult to learn $\vartheta = \prod_A \vartheta$ within L_2-transportation distance ϵ_1 using a sample size independent of n. This is mainly due to the rotationally-invariant nature of L_2, which makes this equivalent to a learning problem in $\mathbb{R}^k$. However, the same is not true for the L_1 distance. Note that we place no assumptions on A, so in order to obtain an estimate $\widetilde{\vartheta}$ with $\mathrm{Tran}_1(\widetilde{\vartheta}, \vartheta) \le \epsilon_1$, we essentially need to ensure that $\mathrm{Tran}_2(\widetilde{\vartheta}, \vartheta) \le \epsilon_1/\sqrt{n}$; however, this would require a sample size depending on n. It is precisely to prevent this $\sqrt{n}$-factor loss that we re-

quire that an L_2-ball in our subspace $\mathrm{Span}(B)$ be close to an L_∞-ball (and hence, an L_1-ball is "nearly spherical"). This ensures that ϑ_B is supported in an L_2-ball of radius $L = O(1/\sqrt{n})$, which makes it possible to learn ϑ_B within Tran_2-distance $\epsilon_1/\sqrt{n}$ with sample size independent of n, since the desired error is $O(L)$. The standard SVD method would typically return the subspace spanned by the first few eigenvectors of A; but this suffers from the same problem as when we use the subspace $\mathrm{Span}(A)$, since there is no guarantee that an L_2-ball in this subspace is close to an L_∞-ball in this subspace.

We now state the main result of this section. We use the following parameters throughout the paper. The polynomial in the definition of C below depends on the specific problems and we will instantiate it later.

$$C = \mathrm{poly}\Big(k, \tfrac{1}{\epsilon}\Big), \quad L = O\Big(\sqrt{\tfrac{k}{n}} \cdot \tfrac{C}{\epsilon}\Big), \quad \epsilon_1 = O\Big(\tfrac{\epsilon^2}{\sqrt{k}C}\Big). \qquad (4)$$

THEOREM 4.1. *Let ϑ be an arbitrary mixture on $\mathrm{Span}(A) \cap \Delta^n$ where $\mathrm{Span}(A)$ is a k-dimensional subspace. We can find a subspace $\mathrm{Span}(B)$ of dimension $h \le k$ in polytime such that:*

(i) $\mathrm{Span}(B)$ *satisfies all properties stated in Lemma 4.5 (see below); and*

(ii) *If we can learn an approximation $\widetilde{\vartheta}_B$, supported on $\mathrm{Span}(B)$, of the projected measure $\vartheta_B = \Pi_{\mathrm{Span}(B)}(\vartheta)$ such that $\mathrm{Tran}_1(\vartheta_B, \widetilde{\vartheta}_B) \le \epsilon_1$ using $N_1(n)$, $N_2(n)$ and $N_K(n)$ 1-, 2-, and K-snapshot samples, then we can learn a mixture $\widetilde{\vartheta}$ such that $\mathrm{Tran}_1(\vartheta, \widetilde{\vartheta}) \le \epsilon$ using $O(N_1(n/\epsilon) + n \log n/\epsilon^3)$, $O(N_2(n/\epsilon) + O(k^4 n^3 \log n/\epsilon^6))$ and $O(N_K(n/\epsilon))$ 1-, 2-, and K-snapshot samples respectively.*

The reduction and its analysis. Let r be the vector encoding the 1-snapshot distribution of ϑ, i.e., $r_i = \int x_i \vartheta(\mathrm{d}x) = \Pr[\text{the 1-snapshot sample is } i]$. We say that the mixture ϑ is *isotropic*, if $r_i \in [1/2n, 2/n]$. Using $O(n \log n)$ 1-snapshot samples, we can get sufficiently accurate estimates of r_is with high probability.

LEMMA 4.2 ([36]). *For any $\sigma > 0$, we can use $O(\frac{1}{\sigma^3} n \log n)$ independent 1-snapshot samples to get $\widetilde{r}_i$ such that, with probability at least $1 - 1/n^2$, for all $i \in [n]$,*

$$\widetilde{r}_i \in (1 \pm \sigma) r_i \ \text{ if } r_i \ge \sigma/2n, \qquad \widetilde{r}_i \le (1+\sigma)\sigma/2n \ \text{ if } r_i < \sigma/2n.$$

Next, we show it is without loss of generality to assume that the given mixture is isotropic, at the expense of a small additive error. The argument essentially follows that of [36], but is simpler.

LEMMA 4.3. *Suppose we can learn with probability $1 - \delta$ an isotropic mixture on $[n]$ within L_1 transportation distance ϵ using $N_1(n), N_2(n)$ and $N_K(n)$ 1-, 2-, and K-snapshot samples respectively. Then we can learn, with probability $1 - O(\delta)$, an arbitrary mixture within L_1 transportation distance 2ϵ using $O(\frac{1}{\sigma^3} n \log n + N_1(n/\sigma)), O(N_2(n/\sigma))$ and $O(N_K(n/\sigma))$ 1-, 2-, and K-snapshot samples respectively, where $\sigma < \epsilon/4$.*

From now on, we assume that the given mixture ϑ is isotropic. Let A be the $n \times n$ symmetric matrix encoding the 2-snapshot distribution of ϑ; i.e., A_{ij} is the probability of obtaining a 2-snapshot (i, j). It is easy to see that $A = \int_{\Delta^n} x x^T \vartheta(\mathrm{d}x)$. Note that the support $\mathrm{Support}(\vartheta)$ of the mixture ϑ is contained in the subspace, $\mathrm{Span}(A)$, spanned by the columns of A. For ease of exposition, we first assume that we know A exactly. This assumption can be dropped via somewhat standard matrix perturbation arguments, which we sketch at the end of this section. Consider the hypercube $\mathcal{H} = [-C/n, C/n]^n$ in $\mathbb{R}^n$ (C only depends on k and ϵ, and is fixed later). We now have all the notation to give a detailed description of the reduction.

REDUCTION 1.

Constructing the basis B. *Input:* Matrix A. *Output:* A basis B satisfying Lemma 4.5.

Consider the centrally symmetric polytope $\mathcal{P} = \mathcal{H} \cap \mathrm{Span}(A)$ and the John ellipsoid $\mathcal{E}$ inscribed in $\mathcal{P}$. It is well known that $\mathcal{E} \subseteq \mathcal{P} \subseteq \sqrt{k}\mathcal{E}$. Suppose the principle axes of $\sqrt{k}\mathcal{E}$ are $\{e_1, \ldots, e_k\}$, sorted in nondecreasing order of their lengths. We choose the orthonormal basis B to be $B = \Big\{ b_i = \frac{e_i}{\|e_i\|_2} : \|e_i\|_2 \ge \frac{\epsilon}{\sqrt{n}} \Big\}$.

Final adjustment. *Input:* Matrix B, $\widetilde{\vartheta}_B$ (which is an approximation of ϑ_B and supported on $\mathrm{Span}(B)$).

Output: The final estimation $\widetilde{\vartheta}$ of the original mixture ϑ.

1. Define the polytope $\mathcal{Q} = \big(\Delta^n + \mathsf{B}_1^n(\epsilon)\big) \cap \mathrm{Span}(B)$. Here $\mathsf{B}_1^n(\epsilon)$ denotes the L_1-ball in $\mathbb{R}^n$ with radius ϵ, and the Minkowski sum $A + B$ of sets A and B is the set $\{a + b \mid a \in A, b \in B\}$. Essentially, $\mathcal{Q}$ is the set of points in $\mathrm{Span}(B)$ with L_1 norm within $[1 - \epsilon, 1 + \epsilon]$.
2. Let $\widetilde{\vartheta}_{\mathcal{Q}} = \Pi_{\mathcal{Q}}(\widetilde{\vartheta}_B)$ be the measure $\widetilde{\vartheta}_B$ projected to $\mathcal{Q}$, i.e., $\widetilde{\vartheta}_{\mathcal{Q}}(S) = \widetilde{\vartheta}_B(\Pi_{\mathcal{Q}}^{-1}(S))$ for any $S \subseteq \mathcal{Q}$.
3. Notice that $\widetilde{\vartheta}_{\mathcal{Q}}$ may not be a valid mixture since some points in $\widetilde{\vartheta}_{\mathcal{Q}}$ may not be in Δ^n. In this final step, we L_1-project $\widetilde{\vartheta}_{\mathcal{Q}}$ back into Δ_n and obtain a valid mixture $\widetilde{\vartheta}$ (i.e., for each point in $\mathcal{Q}$, we map it to its L_1-closest point in Δ^n), which is our final estimation of ϑ.

Lemma 4.4 shows that for large enough C, $\mathcal{H}$ contains $(1 - \epsilon)$ unit of mass of ϑ. Lemma 4.5 proves various properties about $\mathrm{Span}(B)$, which we exploit to prove that the final adjustment procedure returns a good estimate of ϑ.

LEMMA 4.4. *For any $\epsilon > 0$, the following hold.* (i) *Suppose ϑ is a k-spike distribution. For $C \ge 3k/\epsilon$, $\vartheta(\mathcal{H}) \ge 1 - \epsilon$.* (ii) *Suppose ϑ is an arbitrary distribution supported in a k-dimensional subspace. For $C \ge 5k^2/\epsilon$, $\vartheta(\mathcal{H}) \ge 1 - \epsilon$.*

PROOF. For part (i), suppose $\vartheta = \sum_{i=1}^k p_i \delta_{\alpha_i}$ where δ_{α_i} is the Dirac delta at point α_i. We use α_{ij} to denote the jth coordinate of α_i. Since ϑ is isotropic, we know that $\sum_{i=1}^k p_i \alpha_{ij} = r_j \in [1/2n, 2/n]$. So, if $\alpha_{ij} > C/n$ for some j (or equivalently $\alpha_i \notin \mathcal{H}$), we have $p_i \le 2/C$. The lemma thus follows since there can be at most k such points.

To show part (ii), consider the two convex polytopes $\mathcal{P}_s = \mathrm{Span}(A) \cap \frac{1}{k}\mathcal{H}$ and $\mathcal{P} = \mathrm{Span}(A) \cap \mathcal{H}$, where $\frac{1}{k}\mathcal{H} = \big[-\frac{C}{kn}, \frac{C}{kn}\big]^n$. Both $\mathcal{P}_1$ and $\mathcal{P}_2$ are symmetric k-dimensional bodies. By a classical result from convex geometry[2], we can find a linear transformation $\mathcal{K}$ of the unit hypercube $[-1, +1]^k$, such that $\mathcal{K} \subset \mathrm{Span}(A)$ and $\mathcal{P}_s \subseteq \mathcal{K} \subseteq k\mathcal{P}_s = \mathcal{P}$.

Now, we confine ourselves to $\mathrm{Span}(A)$. $\mathcal{K}$ has $2k$ faces of codimension 1. For each such face F, consider the polyhedron

$$\mathcal{C}_F = \{x \mid x = \alpha y, \text{ for some } \alpha \ge 1 \text{ and } y \in F\}.$$

[2]This can be seen either from John's theorem, or the fact that Banach-Mazur distance between any two norms in $\mathbb{R}^k$ is at most k (see, e.g., [39]).

In other words, F separates the cone generated by F into two parts and $\mathcal{C}_F$ is the unbounded part. We claim that $\vartheta(\mathcal{C}_F) \le 2k/C$ for any face F. Consider the normalized vector $r_F = \int_{\mathcal{C}_F} x\,\vartheta(\mathrm{d}x)/\vartheta(\mathcal{C}_F)$. Since r_F is a convex combination of vectors in $\mathcal{C}_F$ and $\mathcal{C}_F$ is convex, r_F is in $\mathcal{C}_F$. Moreover, it is easy to see $\mathcal{P}_s \cap \mathcal{C}_F = \emptyset$. So there must be a coordinate of r_F whose value is larger than C/nk. Since $r = \int x\vartheta(\mathrm{d}x) \ge \vartheta(\mathcal{C}_F) r_F$, we must have $\vartheta(\mathcal{C}_F) \le 2k/C$. All such $\mathcal{C}_F$s together fully cover the region outside $\mathcal{P}$, and there are at most $2k$ such $\mathcal{C}_F$s. So the total mass outside $\mathcal{P}$ is at most $4k^2/C$. □

LEMMA 4.5. *Let* $L = O\left(\sqrt{k/n}\cdot C/\epsilon\right)$. *Let* $\mathcal{P} = \mathrm{Span}(A)\cap\mathcal{H}$. *Let* $v \in \mathrm{Span}(B)$. *Then,* (i) *If* $\|v\|_2 = 1$ *then* $\|v\|_\infty \le L$.
(ii) *If* $\|v\|_1 = 1$ *then* $\frac{1}{\sqrt{n}} \le \|v\|_2 \le L$.
(iii) *If* $x \in \mathbb{R}^n$ *with* $\|x\|_1 = 1$, *then* $\|\Pi_B(x)\|_2 \le L$.
(iv) *For every point* $w \in \mathcal{P}$, $\|w - \Pi_B(w)\|_2 \le \epsilon/\sqrt{n}$.

PROOF. Suppose $|B| = h$. Consider the ellipsoid $\mathcal{E}_B = \sqrt{k}\mathcal{E}\cap\mathrm{Span}(B)$. Clearly, the principle axes of $\mathcal{E}_B$ are $e_1, \ldots, e_h$. Suppose u is an arbitrary point in the boundary of $\mathcal{E}_B$ and $v = u/\|u\|_2$ is a unit vector in $\mathrm{Span}(B)$. Obviously, $\|u\|_\infty \le C\sqrt{k}/n$ (as $u \in \sqrt{k}\mathcal{E} \subseteq \sqrt{k}\mathcal{H}$) and $\|u\|_2 \ge \epsilon/\sqrt{n}$. Hence, $\|v\|_\infty = \|u\|_\infty/\|u\|_2 \le L$, which proves part (i).

Now we show part (ii). The first inequality, $\frac{1}{\sqrt{n}} \le \|v\|_2$, is always true. To see the second inequality, we use the Hölder inequality:

$$\|v\|_2^2 = \langle v, v\rangle \le \|v\|_1\|v\|_\infty = \frac{\|v\|_\infty}{\|v\|_2}\cdot\|v\|_2 \le L\|v\|_2.$$

To prove part (iii), use the Hölder inequality again:

$$\|\Pi_B(x)\|_2 = \frac{\langle x, \Pi_B(x)\rangle}{\|\Pi_B(x)\|_2} \le \frac{\|x\|_1\|\Pi_B(x)\|_\infty}{\|\Pi_B(x)\|_2} \le L.$$

For part (iv), consider an arbitrary point $w \in \mathcal{P} = \mathrm{Span}(A)\cap\mathcal{H}$. We can see that $w \in \sqrt{k}\mathcal{E}$. By the construction of B, any point in $\sqrt{k}\mathcal{E}$ has an L_2 distance at most $\|e_{h+1}\|_2$ from $\mathrm{Span}(B)$, so does w. □

We now prove part (ii) of Theorem 4.1. Let $\widetilde{\vartheta}_B$ supported on $\mathrm{Span}(B)$ be such that $\mathrm{Tran}_1(\vartheta_B, \widetilde{\vartheta}_B) \le \epsilon_1$. Define $\vartheta_{\mathcal{Q}} = \Pi_{\mathcal{Q}}(\vartheta)$ to be the original measure ϑ projected to $\mathcal{Q}$.

LEMMA 4.6. *We have that* $\mathrm{Tran}_1(\vartheta_{\mathcal{Q}}, \vartheta) \le O(\epsilon)$.

PROOF. For any measure μ and subset $S \subset \mathbb{R}^n$, let $\mu|_S$ be the measure A restricted to S. It is easy to see that

$$\mathrm{Tran}_1(\vartheta, \vartheta_{\mathcal{Q}}) \le \mathrm{Tran}_1(\vartheta|_{\mathcal{H}}, \Pi_{\mathcal{Q}}(\vartheta|_{\mathcal{H}})) + \mathrm{Tran}_1(\vartheta|_{\overline{\mathcal{H}}}, \Pi_{\mathcal{Q}}(\vartheta|_{\overline{\mathcal{H}}}))$$

where $\mathcal{H} = [-C/n, C/n]^n$ (the hypercube used in Lemma 4.4). Note that even though the transportation distance is measure in L_1, the projection is with respect to L_2 distance in this lemma. We first bound the term $\mathrm{Tran}_1(\vartheta|_{\overline{\mathcal{H}}}, \Pi_{\mathcal{Q}}(\vartheta|_{\overline{\mathcal{H}}}))$ by coupling every point $p \in \Delta^n$ and $\Pi_{\mathcal{Q}}(p)$ together. By Lemma 4.5 (iv), the L_2 distance from every point in $\mathcal{P} = \mathrm{Span}(A) \cap \Delta^n \cap \mathcal{H}$ is at most $\epsilon/\sqrt{n}$ from $\mathrm{Span}(B)$. Hence, $\|p - \Pi_B(p)\|_1 \le \sqrt{n}\|p - \Pi_B(p)\|_2 \le \epsilon$ and $\|\Pi_B(p)\|_1 \le \|p\|_1 + \|p - \Pi_B(p)\|_1 \le 1 + \epsilon$, which implies $\Pi_{\mathcal{Q}}(p) = \Pi_B(p)$. Thus the first term is at most ϵ.

Now, we bound the second term. For any point $p \in \Delta^n$, it is easy to see the L_1 distance from p to $\Pi_{\mathcal{Q}}(p)$ is at most $2 + \epsilon$. Since the total mass in $\vartheta|_{\overline{\mathcal{H}}}$ is at most ϵ, $\mathrm{Tran}_1(\vartheta|_{\overline{\mathcal{H}}}, \Pi_{\mathcal{Q}}(\vartheta|_{\overline{\mathcal{H}}}))$ is at most $(2+\epsilon)\epsilon < 3\epsilon$. □

LEMMA 4.7. *Let* $\epsilon_1 = O(\epsilon^2/\sqrt{k}C)$. *Let* $\widetilde{\vartheta}_{\mathcal{Q}}$ *be as defined in Reduction 1 and suppose* $\widetilde{\vartheta}_B$ *is such that* $\mathrm{Tran}_1(\vartheta_B, \widetilde{\vartheta}_B) \le \epsilon_1$. *Then, it holds that* $\mathrm{Tran}_1(\vartheta_{\mathcal{Q}}, \widetilde{\vartheta}_{\mathcal{Q}}) \le O(\epsilon)$.

PROOF. First, notice that $\vartheta_{\mathcal{Q}} = \Pi_{\mathcal{Q}}(\vartheta) = \Pi_{\mathcal{Q}}(\Pi_{\mathrm{Span}(B)}(\vartheta)) = \Pi_{\mathcal{Q}}(\vartheta_B)$. So, we have

$$\mathrm{Tran}_2(\vartheta_{\mathcal{Q}}, \widetilde{\vartheta}_{\mathcal{Q}}) = \mathrm{Tran}_2(\Pi_{\mathcal{Q}}(\vartheta_B), \Pi_{\mathcal{Q}}(\widetilde{\vartheta}_B)) \le \mathrm{Tran}_2(\vartheta_B, \widetilde{\vartheta}_B),$$

where the last inequality holds since L_2-projection to a convex set is a contraction and Lemma 2.1 (i). Lemma 4.5 (ii), we have $\mathrm{Tran}_2(\vartheta_B, \widetilde{\vartheta}_B) \le L \cdot \mathrm{Tran}_1(\vartheta_B, \widetilde{\vartheta}_B)$. Therefore,

$$\mathrm{Tran}_1(\vartheta_{\mathcal{Q}}, \widetilde{\vartheta}_{\mathcal{Q}}) \le \sqrt{n}\,\mathrm{Tran}_2(\vartheta_{\mathcal{Q}}, \widetilde{\vartheta}_{\mathcal{Q}}) \le \sqrt{n}\cdot L\cdot\mathrm{Tran}_1(\vartheta_B, \widetilde{\vartheta}_B).$$

Plugging in the value $L = O(\sqrt{k/n}\cdot C/\epsilon)$, we prove the lemma. □

PROOF OF PART (II) OF THEOREM 4.1. By Lemmas 4.6 and 4.7, we have $\mathrm{Tran}_1(\vartheta, \widetilde{\vartheta}_{\mathcal{Q}}) \le \mathrm{Tran}_1(\vartheta, \vartheta_{\mathcal{Q}}) + \mathrm{Tran}_1(\vartheta_{\mathcal{Q}}, \widetilde{\vartheta}_{\mathcal{Q}}) \le O(\epsilon)$. By considering the coupling between all points in $\mathcal{Q}$ and the corresponding points in $\mathrm{Support}(\widetilde{\vartheta})$, we can see that $\widetilde{\vartheta}$ is the probability measure supported in Δ^n that has the closest L_1-transportation distance to $\widetilde{\vartheta}_{\mathcal{Q}}$. Hence, $\mathrm{Tran}_1(\widetilde{\vartheta}, \widetilde{\vartheta}_{\mathcal{Q}}) \le \mathrm{Tran}_1(\vartheta, \widetilde{\vartheta}_{\mathcal{Q}}) \le O(\epsilon)$. We conclude by noting that $\mathrm{Tran}_1(\vartheta, \widetilde{\vartheta}) \le \mathrm{Tran}_1(\vartheta, \widetilde{\vartheta}_{\mathcal{Q}}) + \mathrm{Tran}_1(\widetilde{\vartheta}_{\mathcal{Q}}, \widetilde{\vartheta}) \le O(\epsilon)$. □

***A* is unknown.** We now remove the assumption that A is known. First, we obtain a close approximation of A using $O(k^4n^3\log n/\epsilon^6)$ 2-snapshot samples as follows. We choose a Poisson random variable N_2 with $\mathbb{E}[N_2] = O(k^4n^3\log n/\epsilon^6)$, choose N_2 independent 2-snapshots, and construct a symmetric $n \times n$ matrix $\widetilde{A}$ where $\widetilde{A}_{ii}$ is the frequency of the 2-snapshot (i, i), for all $i \in [n]$, and $\widetilde{A}_{ij} = \widetilde{A}_{ji}$ is half of the total frequency of the 2-snapshots (i, j) and (j, i), for all $i \ne j$.

LEMMA 4.8. *The matrix* $\widetilde{A}$ *obtained above with* $\mathbb{E}[N_2] = O\left(\frac{k^4n^3\log n}{\epsilon^6}\right)$ *satisfies* $\|A - \widetilde{A}\| \le O\left(\frac{\epsilon^3}{k^2n^{3/2}}\right)$.

We find the basis $\widetilde{B}$ as described in Reduction 1, except that we use $\widetilde{A}$ instead of A. Since $\widetilde{B}$ satisfies all properties in Lemma 4.5, the algorithms and analysis in Sections 5 and 6 continue to work. Suppose that we have an estimate $\widetilde{\vartheta}_{\widetilde{B}}$ of $\vartheta_{\widetilde{B}} = \Pi_{\widetilde{B}}(\vartheta)$ such that $\mathrm{Tran}_1(\widetilde{\vartheta}_{\widetilde{B}}, \vartheta_{\widetilde{B}}) \le \epsilon_1$. We project $\widetilde{\vartheta}_{\widetilde{B}}$ to $\widetilde{\mathcal{Q}} = (1+\epsilon)\Delta^n\cap\mathrm{Span}(\widetilde{B})$ to obtain $\widetilde{\vartheta}_{\widetilde{\mathcal{Q}}}$. The same proof as that of Lemma 4.7 shows that $\mathrm{Tran}_1(\vartheta_{\widetilde{\mathcal{Q}}}, \widetilde{\vartheta}_{\widetilde{\mathcal{Q}}}) \le O(\epsilon)$. So the only remaining task is to prove an analogue of Lemma 4.6 showing that $\vartheta_{\widetilde{\mathcal{Q}}}$ is close to the original mixture ϑ.

LEMMA 4.9. *We have that* $\mathrm{Tran}_1(\vartheta_{\widetilde{\mathcal{Q}}}, \vartheta) \le O(\epsilon)$.

5. LEARNING ARBITRARY MIXTURES IN A K-DIMENSIONAL SUBSPACE

Suppose that ϑ is an arbitrary distribution supported on a k-dimensional subspace $\mathrm{Span}(A)$ in $\mathbb{R}^n$. It is known that in order to learn ϑ within transportation distance ϵ, it is necessary to use K-snapshot samples with $K = \Omega(1/\epsilon)$ [36], even in the 1-dimensional case. In this section, we generalize the result to higher dimensions. By the reduction in Theorem 4.1, we only need to specify how to learn a good approximation $\widetilde{\vartheta}_B$ of ϑ_B such that $\mathrm{Tran}_1(\vartheta_B, \widetilde{\vartheta}_B) \le \epsilon_1$. This

can be done as follows. $B = \{b_1, \ldots, b_h\}$ is an $n \times h$ matrix (Recall that B is an orthonormal basis for $\mathrm{Span}(B)$). Let $b'_1, \ldots, b'_n$ be columns of B^T. We use the following parameters in this section: $C = O(k^2/\epsilon)$ as suggested in Lemma 4.4, ϵ_1 and L are as in (4), and

$$\epsilon_2 = \frac{\epsilon_1}{L\sqrt{n}} = \left(\frac{\epsilon}{k}\right)^5, \quad K = O\left(\frac{h}{\epsilon_2^2}\log\frac{h}{\epsilon_2}\right), \quad N = O\left(\frac{1}{\epsilon_2}\right)^h.$$

Suppose we take a K-snapshot sample $\mathsf{s} = \{\ell_1, \ldots, \ell_K\}$ from ϑ, where $\ell_i \in [n]$ for $i = 1, \ldots, K$. Let $\widetilde{\mu}(\mathsf{s}) = \frac{1}{K}\sum_{i=1}^K b'_{\ell_i}$ (which is an h-vector). Suppose we have N K-snapshot samples $\{\mathsf{s}_1, \ldots, \mathsf{s}_N\}$. We define the empirical measure $\widetilde{\mu} = \frac{1}{N}\sum_{i=1}^N \delta(\widetilde{\mu}(\mathsf{s}_i))$, where $\delta()$ is the Dirac delta measure. Our estimation for ϑ_B is $\widetilde{\vartheta}_B = B\widetilde{\mu} = \frac{1}{N}\sum_{i=1}^N \delta(B\widetilde{\mu}(\mathsf{s}_i))$. Note that $\widetilde{\vartheta}_B$ is indeed a discrete measure supported on $\mathbb{R}^n$ as $B\widetilde{\mu}(\mathsf{s}_i)$ is an n-vector. We can also see that $\widetilde{\mu} = B^T\widetilde{\vartheta}_B$ since $B^TB = I$.

Analysis. First, we define μ to be the measure ϑ_B, represented in basis B. Hence, μ is supported over $\mathbb{R}^h$. Formally, $\mu = B^T\vartheta_B = B^T\Pi_B\vartheta = B^TBB^T\vartheta = B^T\vartheta$. Now, we show that $\widetilde{\mu}$ is a good estimation of μ. For this purpose, we introduce an intermediate measure μ_N defined as follows: Suppose the K-snapshot sample s_i is obtained from distribution $s_i \in \mathrm{Span}(A) \cap \Delta^n$. Note that s_i is an n-vector and let $\vartheta_N = \sum_{i=1}^N \delta(s_i)$ and $\mu_N = B^T\vartheta_N$. First, we show μ_N and $\widetilde{\mu}$ are close.

LEMMA 5.1. *Let μ_N and $\widetilde{\mu}$ be defined as above and $K = O(\frac{h}{\epsilon_2^2}\log\frac{h}{\epsilon_2})$. Then, $\mathrm{Tran}_2(\mu_N, \widetilde{\mu}) \le O(\epsilon_2 L)$.*

PROOF. We simply couple $B^Ts_i \in \mathrm{Support}(\mu_N)$ and $\widetilde{\mu}(\mathsf{s}_i) \in \mathrm{Support}(\widetilde{\mu}')$ together. Conditioning on s_i, we can see that $\mathbb{E}[\widetilde{\mu}(\mathsf{s}_i)] = B^Ts_i$. Recall from Lemma 4.5 that the magnitude of every entry of B is at most L. By a standard application of the Chernoff-Hoeffding bound and a union bound over h coordinates, we can see that $\Pr[\|\widetilde{\mu}(\mathsf{s}_i) - B^Ts_i\|_\infty > \epsilon_2 L/\sqrt{h}] < he^{-2\epsilon_2^2 K/h} \le \epsilon_2/2$. Hence, with high probability, for at least $(1-\epsilon_2)N$ samples s_i, we have $\|\widetilde{\mu}(\mathsf{s}_i) - B^Ts_i\|_2 < \epsilon_2 L$. Moreover, $\|\widetilde{\mu}(\mathsf{s}_i) - B^Ts_i\|_2 \le O(L\sqrt{h})$ for all i. So, $\mathrm{Tran}_2(\mu_N, \widetilde{\mu}) \le (1-\epsilon_2)\cdot\epsilon_2 L + \epsilon_2 \cdot O(L\sqrt{h}) \le O(\epsilon_2 L)$. □

LEMMA 5.2. *Let μ and μ_N be defined as above and $N = O(1/\epsilon_2)^h$. Then, with probability at least $1-\epsilon_2$, it holds that $\mathrm{Tran}_2(\mu, \mu_N) \le O(\epsilon_2 L)$.*

PROOF. μ_N is the empirical measure of μ. It is well known that $\mu_N \to \mu$ almost surely in the topology of weak convergence. In particular, the rate of convergence, in terms of transportation distance, can be bounded as follows [2, 43]: for any ϵ_2, for $N > C$ for some large constant C depending only on ϵ_2, with probability at least $1-\epsilon_2$, we have $\mathrm{Tran}_2(\mu_N, \mu) \le O\left(L/N^{1/h}\right)$. Plugging $N = O(1/\epsilon_2)^h$ yields the result. □

Combining Lemmas 5.1 and 5.2, we obtain $\mathrm{Tran}_2(\mu, \widetilde{\mu}) = \mathrm{Tran}_2(B^T\vartheta_B, B^T\widetilde{\vartheta}_B) \le O(\epsilon_2 L)$. Viewing B as an operator from $L_2(\mathbb{R}^h)$ to $L_1(\mathbb{R}^n)$, its operator norm is

$$\|B\|_{2\to1} = \sup_{x\in\mathbb{R}^h}\frac{\|Bx\|_1}{\|x\|_2} = \sup_{x\in\mathbb{R}^h}\frac{\|Bx\|_1}{\|Bx\|_2} \le \sqrt{n}.$$

So by Lemma 2.1, $\mathrm{Tran}_1(\vartheta_B, \widetilde{\vartheta}_B) = \mathrm{Tran}_1(B\mu, B\widetilde{\mu})$, which is at most $\|B\|_{2\to1}\,\mathrm{Tran}_2(\mu, \widetilde{\mu}) \le O(\epsilon_2 L\sqrt{n}) \le \epsilon_1$.

Combining with Theorem 4.1, we obtain the following theorem for learning an arbitrary (even continuous) k-dimensional mixture. The sample size bounds for 1- and 2-snapshots below follow from Lemma 4.2 (taking $\sigma = O(\epsilon)$) and Lemma 4.8.

THEOREM 5.3. *Let ϑ be a mixture supported on $\mathrm{Span}(A) \cap \Delta_n$, where $\mathrm{Span}(A)$ is a k-dimensional subspace. Using $O(n\log n/\epsilon^3)$, $O(k^4n^3\log n/\epsilon^6)$, and $\left(\frac{k}{\epsilon}\right)^{O(k)}$ 1-, 2-, and K-snapshot samples respectively, where $K = \widetilde{O}(k^{11}/\epsilon^{10})$, we can obtain, with probability 0.99, a mixture $\widehat{\vartheta}$ such that $\mathrm{Tran}_1(\widetilde{\vartheta}, \vartheta) \le O(\epsilon)$*

6. LEARNING K-SPIKE MIXTURES ON Δ_N

In this section, we consider the setting where ϑ is a k-spike distribution on Δ_n, that is, ϑ is supported on k points in Δ_n. This setting was also considered in [36] but unlike the results therein, our sample size bounds *only depend on n and k* and *not* on any "width" parameters of ϑ (e.g., the least weight of a mixture constituent, or the distance between two spikes). We use K-snapshot samples only for $K = 2k-1$ in this section, which is known to be necessary [36].

The high level idea of our algorithm is as follows. Again, given the reduction of Section 4, we only need to provide an algorithm for learning a good approximation $\widetilde{\vartheta}_B$ for the projected measure $\vartheta_B := \Pi_{\mathrm{Span}(B)}(\vartheta)$. More specifically, we need $\mathrm{Tran}_1(\widetilde{\vartheta}_B, \vartheta_B) \le \epsilon_1$. For this purpose, we pick a fine net of directions in $\mathrm{Span}(B)$ and learn the 1-dimensional projected measures on these directions. Then we use the 1-dimensional projected measures to reconstruct $\Pi_{\mathrm{Span}(B)}\vartheta$. The reconstruction can be done by a linear program that is similar to LP_1 in Section 3.1. The most crucial and technically challenging part is to show that if the 1D-projections of two measures are close (in Tran), then the two measures must be close as well. To do this, we leverage Yudin's theorem (Theorem 2.6), which shows that any 1-Lip-function f in $\mathsf{B}_2^h(1)$ admits a good approximation in terms of certain 1D-functions with bounded Lipschitz constant. Since the 1D-projections of the two measures are close, the Kantorovich-Rubinstein theorem implies that the RHS of (3) is small for these 1D functions, and hence that the RHS of (3) is small for f. This implies (again by (3)) that the two measures are close in Tran. We defer the details of the learning algorithm and the proof of the following theorem to the full version of the paper.

THEOREM 6.1. *Let ϑ be an arbitrary k-spike mixture in Δ_n. Using $O(n\log n/\epsilon^3)$, $O(k^4n^3\log n/\epsilon^6)$, and $(k/\epsilon)^{O(k^2)}$ 1- and 2- and $(2k-1)$-snapshot samples respectively, we can obtain, with probability 0.99, a mixture $\widehat{\vartheta}$ such that $\mathrm{Tran}_1(\widetilde{\vartheta}, \vartheta) \le O(\epsilon)$.*

7. REFERENCES

[1] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proc. 18th Ann. Conf. on Learning Theory*, pages 458–469, June 2005.

[2] Kenneth S Alexander. Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Prob.*, 12:1041–1067, 1984.

[3] 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.

[4] 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.

[5] Anima Anandkumar, Yi-kai Liu, Daniel J Hsu, Dean P Foster, and Sham M Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

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

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

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

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

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

[11] Jean Bourgain, Joram Lindenstrauss, and V Milman. Approximation of zonoids by zonotopes. *Acta mathematica*, 162(1):73–141, 1989.

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

[13] 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.

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

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

[16] 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.

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

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

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

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

[21] Richard M Dudley. A course on empirical processes. In *Ecole d'Eté de Probabilités de Saint-Flour XII-1982*, pages 1–142. Springer, 1984.

[22] Richard M Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

[23] 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.

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

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

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

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

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

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

[30] 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.

[31] Jon Kleinberg and Mark Sandler. Using mixture models for collaborative filtering. *JCSS*, 74:49–69, 2008.

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

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

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

[35] Abedallah Rababah. Transformation of chebyshev–bernstein polynomial basis. *Comput. Methods Appl. Math.*, 3(4):608–622, 2003.

[36] Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *Proc. 5th ITCS*, pages 207–224, 2014.

[37] Theodore J Rivlin. *An introduction to the approximation of functions*. Courier Dover Publications, 2003.

[38] GW Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. 1990.

[39] Nicole Tomczak-Jaegermann. *Banach-Mazur distances and finite-dimensional operator ideals*, volume 38. Longman Scientific & Technical Harlow, 1989.

[40] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. *JCSS*, 68:841–860, 2004.

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

[42] V. A. Yudin. The multidimensional Jackson theorem. *Mathematical Notes*, 20(3):801–804, 1976.

[43] JE Yukich. Optimal matching and empirical measures. *Proceedings of the AMS*, 107(4):1051–1059, 1989.