Inference from Sparse Sampling

(Extended Abstract)

Yuval Rabani* Leonard J. Schulman† Chaitanya Swamy‡

May 20, 2008

Abstract

In some statistical learning applications, the individual objects that are encountered are complex, behave probabilistically, and are encountered too briefly for the learner to form more than a fragmentary and unreliable record of the properties of each individual. In some cases, only previously collected sporadic survey data is available to the learner. In other cases, studying an object extensively may alter the object or may encounter the object's—perhaps a person's—resistance.

Motivated by such considerations, we initiate the study of learning mixture models from sparse samples of their component distributions. We design and analyze learning algorithms for two fundamental mixture models that are loosely motivated by data mining and market analysis applications, and that cannot be learnt by sampling from the mixture distribution. The first model we consider is a mixture of biased coins. The input sample is generated by choosing random coins from the mixture and tossing each of them K times (where K is too small to learn the bias of any single coin accurately). The second model, which is somewhat closer to the motivating applications but requires the first for its solution, is a generalization where the coins are replaced by biased n-faceted dice, for n possibly much larger than K.

*rabani@cs.technion.ac.il. Computer Science Department, Technion — Israel Institute of Technology, Haifa 32000, Israel. Part of this work was done while visiting UCLA and Caltech. Supported in part by ISF 52/03, BSF 2002282, and the Fund for the Promotion of Research at the Technion.

†schulman@caltech.edu. Caltech, Pasadena, CA 91125. Supported in part by NSF CCF-0515342, NSA H98230-06-1-0074, and NSF ITR CCR-0326554.

‡cswamy@math.uwaterloo.ca. Dept. of Combinatorics and Optimization, Univ. Waterloo, Waterloo, ON N2L 3G1. Supported in part by NSERC grant 32760-06. Part of this work was done while the author was a postdoctoral scholar at Caltech, Pasadena CA 91125.

1 Introduction

Problem statement. A statistical *mixture model* is a probability distribution on items (or "components") which are themselves probability distributions; with the restriction that the component distributions take values in a common space of outcomes. A mixture model generates a *mixture distribution* in an obvious way: first choose at random a component distribution from the mixture, then draw one sample at random from the component distribution. In recent years theoreticians have made remarkable progress in forming a computational theory of learning statistical mixture models from a sequence of independent samples of their mixture distribution. The fundamental mixture models investigated include mixtures of Gaussians [16, 17, 4, 39, 27, 1, 22], mixtures of discrete product distributions [29, 23, 13, 21, 10], and similar models [13, 6, 35, 27, 15]. These papers study the problem of learning from sequences of points in $\mathbb{R}^n$ or in $\{0,1\}^n$ that were generated by independent samples from an unknown mixture distribution. A typical PAC-style learning problem is to design an algorithm that takes a sequence of points generated in this manner and computes with high probability a mixture model whose mixture distribution is statistically close to that of the true model.

A mixture model can generate sample sequences other than the obvious sequence of independent samples from the mixture distribution. For instance, each time a component is selected, several samples rather than just one may be drawn from it; the resulting sequence has very different statistics than those of the mixture distribution. It can therefore be desirable, and is a fundamental algorithmic question in statistics, to learn the mixture model itself, not merely the mixture distribution which it defines. Yet there are mixture models that simply cannot be learnt from the mixture distribution they generate. For example, consider a mixture of Bernoulli trials with success probabilities $p_1, p_2, \ldots, p_k$ and equal weight in the mixture. A sequence generated by the mixture distribution is statistically indistinguishable from a sequence generated by a single Bernoulli trial with success probability $\frac{1}{k}\sum_{i=1}^{k} p_i$. Thus, a natural question that arises in this context is the power of correlation: what can be learnt if each time a component distribution is chosen from the mixture, not one but a few samples are drawn from that distribution? [1]

In this paper, we initiate the study of this question. We examine two fundamental classes of mixture models, in each of which little can be learnt about the mixture model from a sequence of independent samples from the mixture distribution. We analyze the effect of obtaining repeated samples from each independently chosen component of the mixture, and design learning algorithms for these mixture models.

The first of these two models is the above-mentioned mixture of Bernoulli trials. In other words, the model is a mixture of coins of various biases. The available sequence of samples is generated by repeatedly selecting a coin at random (the coins are indistinguishable in appearance) and tossing it a few times before a fresh coin is selected. Our goal is to learn the mixture model that generated the input sequence. We allow sample sequences generated by any probability distribution on coin biases, in particular continuous distributions.

One benefit of learning the mixture model is that we may find out, for example, that the coins are clustered into just a few distinct types. Such a discovery, aside from providing a concise representation for the model, may also help us to classify individual coins as they now need to be tested against a small number of hypotheses, rather than against a continuum of possibilities.

As mentioned above, if each coin in the sample is tossed just once, then the only thing we can infer from the sample is the bias of the mixture distribution. On the other hand, if each coin in the sample is tossed as many times as we wish, we can estimate its bias with high accuracy, and therefore we can easily learn the mixture model. Thus, the intriguing question is how to handle the intermediate case, when each coin is tossed only a few times, making it statistically impossible to learn accurately the bias of any single coin in the sample, yet potentially feasible to aggregate data from many samples to infer much about the mixture model. This is the central topic of our paper. We refer to the key parameter, the number of times K that each coin is tossed, as the

[1] This question makes sense even in the case of mixture models that can in principle be learnt from their mixture distribution, e.g., mixtures of Gaussians. Two very close Gaussians are hard to distinguish, but with repeated sampling, the variance of each is reduced, effectively separating them and greatly reducing the overall sample complexity required to learn the parameters of the mixture model.

sampling aperture available to the learning algorithm.[2]

Our second model is adopted from the collaborative filtering mixture model of Kleinberg and Sandler [31, 32].[3] This second model generalizes the coins model, replacing Bernoulli trials by n-way trials. One way to visualize the model is to think of a mixture of biased n-faceted dice (the faces of each die are numbered 1 through n). The learning algorithm is given a sequence of dice sampled from the mixture, each rolled K times (the sampling aperture). Analogously to the coins model, if $K = 1$ then we can only learn the face probabilities in the mixture distribution and not the mixture model, while if K is very large (at least n) we can learn accurately the face probabilities of each individual die in the sample, and thus learn the mixture model with ease. So the interesting case is when K is somewhere in between these extremes.

Motivation. Our study was loosely motivated by market analysis scenarios and other data mining applications. We wish to learn the global properties of a population of individuals, each of which exhibits probabilistic behavior. We only have a small amount of data on each individual—far less data than is necessary in order to describe adequately that individual. The reasons for this deficiency may vary. Collecting additional data may be expensive, it may affect the individual's behavior, it may encounter the individual's resistance, because only previously collected sporadic data is available for study. In spite of this limitation we wish to build up an accurate model of the entire population, by piecing together the fragmentary data from the individuals we have encountered.

Consider, for example, a survey intended to study movie viewing patterns. When people are asked for their movie preferences, they are unlikely to respond with a ranking of all movies ever produced. Instead, one can realistically expect to get from each person surveyed a very short and fairly random list of favorites that happen to cross this person's mind when asked. Movies that are ranked high on the hypothetical complete list are more likely to be mentioned than movies ranked low on that list. Thus, while we cannot hope to study the preferences of a single person thoroughly, we may still be able to aggregate superficial data from many people to detect overall trends and to segment the population into a number of recurring types. Our dice model is a simplistic model for scenarios such as the movie viewing survey (see [31, 32]): the dice are people, the dice faces are movies, and K is the number of answers given by each person. (Our coin model is, of course, motivated by our reduction from the dice model. But as a basic question in machine learning it is also interesting in its own regard.)

There are various ways in which construction of a statistical mixture model for the population can be useful. The model can be used as a Bayesian prior. After a customer has ordered a few movies (or other products), the merchant can make a posteriori Bayesian inference about the preferences of this customer, and shape advertising or recommendations to the customer. (In this context our algorithms can be thought of as a preprocessing phase.) More generally, market research can influence what products the merchant offers. Since market survey costs may vary with their thoroughness, analyzing the limits of market predictions may help in optimizing spending on market studies. But in this paper we stick with the statistical fundamentals, and acknowledge the gap between our work and concrete applications.

Performance measures. The natural metric in which to measure the output quality in our setting is the transportation cost metric [25, 40].[4] Informally, a metric space (U, δ) generically induces a transportation metric $\mathrm{Tran}_{U,\delta}$ on the set of probability distributions on U.[5] The distance $\mathrm{Tran}_{U,\delta}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ between two probability distributions is informally the minimum cost of a flow that redistributes probability mass from $\boldsymbol{\theta_1}$ to $\boldsymbol{\theta_2}$.

A biased coin corresponds to a point $p \in [0, 1]$, where p is the "heads" probability of this coin. Thus, a mixture of coins is a probability measure ϑ on $U = [0, 1]$. We will endow U with the standard metric $\delta(x, y) = |x - y|$. For two mixtures ϑ, ϑ', the distance $\mathrm{Tran}(\vartheta, \vartheta')$ is simply the minimum cost of a matching of the coins in ϑ to

[2] For simplicity, we assume that the sampling aperture is uniform for all sampled objects.

[3] They, in turn, relate their model to that of Hofmann and Puzicha [26].

[4] The transportation metric is also called earthmover, Wasserstein, or exponent-1 Monge-Kantorovich distance.

[5] We will use the simpler notation Tran whenever the underlying metric space (U, δ) is clear from the context.

those in ϑ', where the cost of matching two coins is the difference in their "heads" probabilities. Similarly, a die is a vector $p \in \Delta_n$ denoting its face distribution. A mixture of dice is a probability measure $\boldsymbol{\theta}$ on $U = \Delta_n$. We will endow U with the total variation distance $\delta(x, y) = \frac{1}{2}\|x - y\|_1$. To interpret $\mathrm{Tran}(\boldsymbol{\theta}, \boldsymbol{\theta'})$, think of a min-cost matching of the dice in $\boldsymbol{\theta}$ to those in $\boldsymbol{\theta'}$, where the cost of matching a pair of dice is the total variation distance between their face distributions.

Our results. We first analyze mixtures containing a finite number k of coin types (which we call a k-spike mixture). We prove that in this case the asymptotic statistics of sampling aperture $K = 2k - 1$ characterize the mixture uniquely; this leads to the design of a learning algorithm for the problem. We prove the following theorem.

Theorem 1.1. *There exists a polynomial time algorithm that gets as input $k \in \mathbb{N}$ and a sample of $m \geq (k/W)^{O(k)}$ coins from a k-spike mixture ϑ, each tossed $K = 2k - 1$ times, and outputs a k-spike mixture $\tilde{\vartheta}$, such that $\mathrm{Tran}(\vartheta, \tilde{\vartheta}) \leq W$ with high probability.*

Next we prove that any two coin mixtures with identical asymptotic statistics of sampling aperture K differ by a transportation distance of $O(1/\sqrt{K})$. This applies also to mixtures containing a continuum of coin types, and leads to the design of a learning algorithm for arbitrary coin mixtures. We prove the following theorem.

Theorem 1.2. *There exists a polynomial time algorithm that gets as input a sample of $m \geq \exp K^{O(1)}$ coins from a mixture ϑ, each tossed K times, and outputs a mixture $\tilde{\vartheta}$, such that $\mathrm{Tran}(\vartheta, \tilde{\vartheta}) = O(1/\sqrt{K})$ with high probability.*[6]

The bound on $\mathrm{Tran}(\vartheta, \tilde{\vartheta})$ in Theorem 1.2 is close to best possible, even information-theoretically: there are two mixtures ϑ and $\tilde{\vartheta}$ that produce *identical* statistics of sampling aperture K, yet have $\mathrm{Tran}(\vartheta, \tilde{\vartheta}) = \Omega(1/K)$.[7] We do not know if the exponential dependence of the sample size m on K in Theorems 1.1 and 1.2 is necessary, though we conjecture that this is indeed the case. Notice that Theorems 1.1 and 1.2 are incomparable. The former applies to a special case of the latter, but provides stronger guarantees in the sense that the error does not depend on the sampling aperture, only on the sample size.[8]

Our main result deals with the dice problem. We do not know how to handle arbitrary mixtures $\boldsymbol{\theta}$ (i.e., arbitrary distributions on the simplex Δ_n), and we suspect that they may require prohibitive sample sizes (in terms of n). In this extended abstract we only deal with mixtures $\boldsymbol{\theta}$ of two distributions $p, q \in \Delta_n$ and, more generally, mixtures containing any number (even a continuum) of distributions of the form $ap + (1 - a)q$. We believe that our approach is likely to extend to mixtures containing any finite number of arbitrary distributions in Δ_n. We show that this data model reduces to the coins data model. More specifically, given two distributions $p, q \in \Delta_n$, we can map die rolls to coin tosses as follows. If the die shows a face i for which $q_i > p_i$, then the coin shows "heads," and if the die shows a face i for which $q_i < p_i$, then the coin shows "tails" (we ignore all other rolls). This maps the set distributions $r \in \Delta_n$ with $\Pr_r[i : \ p_i \neq q_i] > 0$ to $[0, 1]$, and in particular it maps the set of distributions of the form $ap + (1 - a)q$ isometrically into $[0, 1]$. We denote this mapping by $\psi_{p,q}$. Notice that points in $[0, 1]$ can be mapped inversely to distributions of the form $ap + (1 - a)q$. We denote this mapping by $\psi_{p,q}^{-1}$. We show the following reduction.

Theorem 1.3. *There is a polynomial time algorithm that gets as input a sample of $m \geq (1/W)^{O(1)} \cdot n \cdot \log^{O(1)} n$ dice from a mixture $\boldsymbol{\theta}$ containing distributions of the form $ap + (1 - a)q$ (for two unknown distributions $p, q \in$*

[6] The Weierstrass approximation theorem ensures that polynomial approximations converge to continuous functions on bounded intervals in L_∞ (implying that under mild conditions, if all the moments of ϑ are known, then it is uniquely defined). As part of our proof of Theorem 1.2, we prove an upper bound on the rate of convergence in terms of the transportation norm (rather than L_∞).

[7] The proof is omitted for lack of space.

[8] Also notice that the k-spike algorithm cannot be used to solve the general case, despite the fact that every distribution can be approximated in transportation cost by a k-spike distribution. The reason is that we sample the true distribution rather than the approximation, so the empirical statistics do not converge to those of a k-spike distribution, and the Theorem 1.1 algorithm may fail.

Δ_m*), each die rolled twice, and outputs two distributions* $\tilde{p}, \tilde{q} \in \Delta_n$ *such that for every probability measure* ϑ *on* $[0,1]$, $\mathrm{Tran}(\boldsymbol{\theta}, \psi_{\tilde{p},\tilde{q}}^{-1}(\vartheta)) \leq \mathrm{Tran}(\psi_{\tilde{p},\tilde{q}}(\boldsymbol{\theta}), \vartheta) + W$.

This gives polynomial time learning algorithms for such mixtures $\boldsymbol{\theta}$, using a sample size of $n \log^{O(1)} n$ (assuming that K and the desired accuracy are fixed), as follows. First compute $\tilde{p}, \tilde{q}$ using Theorem 1.3. Next learn $\psi_{\tilde{p},\tilde{q}}(\boldsymbol{\theta})$ using Theorem 1.1 (if applicable) or Theorem 1.2. Finally, map the result back to Δ_n using $\psi_{\tilde{p},\tilde{q}}^{-1}$. Notice that if the expected face probabilities in the mixture are all $\Theta\left(\frac{1}{n}\right)$, then a constant fraction of the faces will be missed altogether by sampling $O(n/K)$ dice. Thus, we cannot hope to improve the required sample size by more than a $\log^{O(1)} n$ factor.

Organization. The rest of this paper is organized as follows. Section 2 deals with the coins model. In subsection 2.1 we deal with the k-spike case, and in subsection 2.2 we deal with the general case. Section 3 deals with the dice model. We provide some intuition and informal discussion of our algorithms and proof techniques in the beginning of each section. Notation and definitions are also presented there.

2 The Coins Problem

In this section we present our algorithms for the coins problem. A coin is identified with a point x in the interval $J = [0,1]$, the probability of the coin showing "heads." (But in Section 2.2 for convenience we will switch to $J = [-1,+1]$, with "heads" probability $p_x = (1-x)/2$.) The (unknown) mixture model is a probability distribution ϑ on J.

Consider the moment map $\mu' : [0,1] \to \mathbb{R}^K$ that maps $x \in [0,1]$ to the vector $(x, x^2, x^3, \ldots, x^K)$. The image of $[0,1]$ under μ' is called the *moment curve*. The first K polynomial moments of a probability measure ϑ on $[0,1]$ are $\mathrm{E}[\mu'(x)_1], \mathrm{E}[\mu'(x)_2], \ldots, \mathrm{E}[\mu'(x)_K]$, where $x \sim \vartheta$. We will extend the range of μ' to probability measures on $[0,1]$ and write $\mu'(\vartheta) = (\mathrm{E}[\mu'(x)_1], \mathrm{E}[\mu'(x)_2], \ldots, \mathrm{E}[\mu'(x)_K])$. Let X be the random variable denoting the number of times a coin drawn from ϑ and tossed K times shows "heads." We can compute $\mu'(\vartheta)$ if we know the distribution of X, given by the following equations. For every $i \in \{0, 1, \ldots, K\}$, $\Pr[X = i] = \int_J \binom{K}{i} x^i (1-x)^{K-i} d\vartheta(x)$. The latter probabilities can be estimated empirically to within any desired accuracy given a sufficiently large sample from ϑ (where each sampled coin is tossed K times). In particular, let $\widetilde{\mathsf{freq}}_j$ denote the fraction of sampled coins that showed "heads" exactly j times. Our algorithms use these empirical frequencies to compute a close approximation $\tilde{\vartheta}$ of ϑ.

It will be convenient to rescale the vector of empirical statistics as follows: $\tilde{f} = \left(\widetilde{\mathsf{freq}}_0, \frac{\widetilde{\mathsf{freq}}_1}{K}, \frac{\widetilde{\mathsf{freq}}_2}{\binom{K}{2}}, \ldots, \widetilde{\mathsf{freq}}_K\right)$. In the limit of large sample size this vector converges to the *moment vector* $f = (f_0, \ldots, f_K) = \left(\mathsf{freq}_0, \frac{\mathsf{freq}_1}{K}, \frac{\mathsf{freq}_2}{\binom{K}{2}}, \ldots, \mathsf{freq}_K\right)$; we define this vector to be $\mu(\vartheta)$, and observe that μ is a linear mapping on the space of signed measures on J.

We need to be more rigorous about transportation cost. Say that a signed measure ϑ on J is bounded if $\vartheta(S)$ exists and is finite for every measurable $S \subseteq J$. Define the trace of a bounded signed measure on J to be $\vartheta(J)$.[9] If ϑ_1, ϑ_2 are two signed measures on J of equal trace, a *transport* of ϑ_1 to ϑ_2 is a measure ν on $J \times J$ such that for all measurable sets $I \subseteq J$, $\vartheta_1(I) = \nu(I, J)$ and $\vartheta_2(I) = \nu(J, I)$. The *transportation distance* $\mathrm{Tran}(\vartheta_1, \vartheta_2)$ is defined to be $\inf \int |x - y| \, d\nu$, where the infimum is taken over all transports ν of ϑ_1 to ϑ_2. Observe that there is a separate metric space for each value of the trace. (Transportation cost for distributions of finite support has a classic LP formulation, see Appendix C.) Define $\bar{\mathcal{T}}_0(J)$ to be the space of trace-0 bounded signed measures on J, with the norm $\|\vartheta\|_{\mathrm{Tran}} = \mathrm{Tran}(\vartheta, \mathbf{0})$ where $\mathbf{0}$ is the signed measure assigning measure 0 to every set.

[9] For background see Doob [19] §IX; we use slightly modified terminology.

2.1 Proof of Theorem 1.1: learning k-spike mixture models using aperture $2k-1$

To prove Theorem 1.1, we show how to invert the moment map μ', and we analyze the effect of the sampling noise on the inversion. We show in particular that a k-spike mixture ϑ is uniquely determined by its first $2k-1$ polynomial moments.[10] The proof relies crucially on showing that the moment curve is highly convex, in the sense that given two distant sets of points on the curve with a total of at most $K+1$ points, any weighted combination of the points of one set is distant from any combination of the other set.

A k-spike mixture ϑ can be represented as $\vartheta = \sum_{i=1}^{k} \vartheta_i \delta_{\alpha_i}$ where $\vartheta_i \geq 0$, $\sum \vartheta_i = 1$, and δ_{α_i} is the Dirac measure that puts weight 1 on the point $\alpha_i \in [0,1]$, corresponding to a "coin type" having probability α_i of showing "heads." We use the notation $\vartheta = (\bar{\vartheta}, \alpha)$, where $\bar{\vartheta} = (\vartheta_1, \ldots, \vartheta_k)$ and $\alpha = (\alpha_1, \ldots, \alpha_k)$. Our goal is to infer the weights ϑ_i and the points α_i with small error and thus approximately reconstruct ϑ.

For a vector $x = (x_1, \ldots, x_\ell)$ (with all $0 < x_i < 1$) and for a positive integer b, let $V_b(x)$ be the $\ell \times b$ matrix $(V_b(x))_{ij} = x_i^j$ (with $1 \leq i \leq \ell$ and $0 \leq j \leq b-1$), and let $A_b(x)$ be the $\ell \times b$ matrix $(A_b(x))_{ij} = (1-x_i)^{b-1-j} x_i^j$ (with $1 \leq i \leq \ell$ and $0 \leq j \leq b-1$). Let P be the $2k \times 2k$ lower triangular "Pascal" matrix: for $0 \leq j \leq 2k-1$ and $j+1 \leq i \leq 2k$, $P_{ij} = \binom{2k-j-1}{i-j-1}$. Then $V_{2k}(\alpha) = A_{2k}(\alpha)P$. Note that the moment vector $f_j = \mathsf{freq}_j / \binom{2k-1}{j}$ (for $0 \leq j \leq 2k-1$) is given by the linear transformation $f = \bar{\vartheta} A_{2k}(\alpha)$, It will be convenient in this section to also work with the vector g of "standard" moments (those w.r.t. the basis x^j), $g = fP = \bar{\vartheta} V_{2k}(\alpha)$.

We show how to compute a k-spike distribution $\tilde{\vartheta} = (\tilde{\bar{\vartheta}}, \tilde{\alpha})$ from the empirical (scaled) moments vector $\tilde{f}$, such that $\mathrm{Tran}(\vartheta, \tilde{\vartheta})$ is small. To give some intuition, suppose at first that we know the true moments vector $g = fP = \bar{\vartheta} V_{2k}(\alpha)$. Observe that there is a common vector $\lambda = (\lambda_0, \ldots, \lambda_k)^T$ 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 $\Lambda_{ij} = \lambda_{i-j}$ (with the understanding $\lambda_\ell = 0$ for $\ell \notin \{0, \ldots, k\}$), $V_{2k}(\alpha)\Lambda = 0$. Thus $g\Lambda = \bar{\vartheta} V_{2k}(\alpha)\Lambda = 0$. Overtly this is a system of $2k$ equations but we eliminate the redundancy in Λ by forming the $k \times (k+1)$ matrix $G = G(g)$ defined by $G_{ij} = g_{i+j}$ for $i = 0, \ldots, k-1$ and $j = 0, \ldots, k$; then solving 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 $\big(V_{2k}(\alpha)\Lambda(\lambda)\big)_{i,1} = \sum_{\ell=0}^{k} \lambda_\ell {\alpha_i}^\ell = 0$. 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 values, we can compute $\bar{\vartheta}$ by solving for y the system of linear equations $yV_{2k}(\alpha) = g$. Since we only have the empirical vector $\tilde{f}$ and not f, not all the steps above may be well defined or yield meaningful values. It is also necessary to control the error that results due to the difference between f and $\tilde{f}$, Put $\tilde{g} = \tilde{f}P$. We assume that $\|\tilde{g} - g\|_2 \leq \xi$, where ξ is a parameter we will fix later. The learning algorithm is as follows:

(1) We first solve the minimization problem:

$$\text{minimize} \quad \|x\|_1 \quad \text{subject to} \quad \|G(\tilde{g})x\|_1 \leq 2^k \xi, \quad x_k = 1 \tag{P}$$

to obtain a solution $\tilde{\lambda}$. Note that this minimization problem can be encoded as a linear program. Observe that since $G(\tilde{g})$ has $k+1$ columns and k rows, there is always a feasible solution.

(2) 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 ζ be the smallest separation between distinct α_i-s. Let $\epsilon = \frac{4}{\zeta}((k+1)\xi)^{1/k}$. First we compute $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ values such that $|\hat{\alpha}_i - \bar{\alpha}_i| \leq \epsilon$ for every i in time $\mathrm{poly}\big(\log(\frac{1}{\epsilon})\big)$ using Pan's algorithm [36, Theorem 1.1] [11] We now set $\tilde{\alpha}_i = \mathrm{Re}(\hat{\alpha}_i)$ if $\mathrm{Re}(\hat{\alpha}_i) \in [0,1]$; $\tilde{\alpha}_i = 0$ if $\mathrm{Re}(\hat{\alpha}_i) < 0$; and $\tilde{\alpha}_i = 1$ if $\mathrm{Re}(\hat{\alpha}_i) > 1$.

[10] In other words, we show that the moment map compresses sparse non-negative signals to dimension about twice their support size; moreover we show that this compression is relatively insensitive to noise.

[11] 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 [36] shows that this is essentially without loss of generality.

(3) Finally, we find $\overline{\tilde{\vartheta}}$ by finding the row-vector $y \in [0,1]^k$ that minimizes $\|yV_{2k}(\tilde{\alpha}) - \tilde{g}\|_2$ subject to $\|y\|_1 = 1$. Notice that this is a convex program.

We now proceed with analyzing our algorithm. Theorem 2.1 establishes the information-theoretic component of Theorem 1.1, namely that if two distributions $\vartheta = (\bar{\vartheta}, \alpha)$ and $\tilde{\vartheta} = (\overline{\tilde{\vartheta}}, \tilde{\alpha})$ are distant in transportation distance, then $\bar{\vartheta}V_{2k}(\alpha)$ and $\overline{\tilde{\vartheta}}V_{2k}(\tilde{\alpha})$ are distant in Euclidean distance. Subsequently Proposition 2.5 shows that our reconstruction algorithm efficiently reconstructs $\overline{\tilde{\vartheta}} = (\tilde{\vartheta}_1, \ldots, \tilde{\vartheta}_k)$ and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_k)$ such that $\|\overline{\tilde{\vartheta}}V_{2k}(\tilde{\alpha}) - \tilde{g}\|_2$ is small. Thus, by choosing a large enough sample size so that $\|\tilde{g} - g\|_2$ is sufficiently small, we obtain that $\|\bar{\vartheta}V_{2k}(\alpha) - \overline{\tilde{\vartheta}}V_{2k}(\tilde{\alpha})\|_2$ is small; by Theorem 2.1 this implies that the distribution $\tilde{\vartheta} = (\overline{\tilde{\vartheta}}, \tilde{\alpha})$ is close in transportation distance to the true distribution $\vartheta = (\bar{\vartheta}, \alpha)$.

Theorem 2.1. $\|g - \overline{\tilde{\vartheta}}V_{2k}(\tilde{\alpha})\|_2 \geq \frac{1}{(2k-1)^2 2^{8k-5}} \cdot \Big(\mathrm{Tran}(\vartheta, \tilde{\vartheta})\Big)^{4k-2}$.

Proof. Consider the difference $\boldsymbol{\vartheta} = \vartheta - \tilde{\vartheta}$ (a bounded signed measure), which we also write as $\boldsymbol{\vartheta} = \sum_{i=1}^{2k} \overline{\boldsymbol{\vartheta}}_i \delta_{\boldsymbol{\alpha}_i}$, where $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{2k}\} = \{\alpha_1, \ldots, \alpha_k\} \cup \{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_k\}$ and $\boldsymbol{\alpha}_1 < \ldots < \boldsymbol{\alpha}_{2k}$. (The assumption that the $\boldsymbol{\alpha}_i$'s are distinct is only for notational convenience.) Observe that $\overline{\boldsymbol{\vartheta}}_1, \ldots, \overline{\boldsymbol{\vartheta}}_{2k}$ may be positive or negative. Let $\overline{\boldsymbol{\vartheta}} \in \mathbb{R}^{2k}$ be the row vector $(\overline{\boldsymbol{\vartheta}}_1, \ldots, \overline{\boldsymbol{\vartheta}}_{2k})$. Let $\eta = \|\boldsymbol{\vartheta}\|_{\mathrm{Tran}}$ and let $y = \overline{\boldsymbol{\vartheta}}V_{2k}(\boldsymbol{\alpha})$. So what we need to show is that $\|y\|_2 \geq \frac{1}{(2k-1)^2 2^{8k-5}} \cdot \eta^{4k-2}$.

There is an $1 \leq \ell < 2k$ such that $\left|\sum_1^\ell \overline{\boldsymbol{\vartheta}}_i\right| \cdot (\boldsymbol{\alpha}_{\ell+1} - \boldsymbol{\alpha}_\ell) \geq \eta/(2k-1)$. Let $\delta = \sum_1^\ell \overline{\boldsymbol{\vartheta}}_i$; without loss of generality $\delta \geq 0$, and note that $\delta \leq 1$. Let $s = \boldsymbol{\alpha}_{\ell+1} - \boldsymbol{\alpha}_\ell$, so $(2k-1)\delta s \geq \eta$. (Of course also $\eta \geq \delta s$.)

Denote row i of a matrix M by M_{i*} and column j by M_{*j}. A vector y minimizing $\|y\|_2 = \|\overline{\boldsymbol{\vartheta}}V_{2k}(\boldsymbol{\alpha})\|_2$ subject to the list $\boldsymbol{\alpha}$ and the value of ℓ, must be orthogonal to $V_{2k}(\boldsymbol{\alpha})_{i*} - V_{2k}(\boldsymbol{\alpha})_{i'*}$ if $1 \leq i < i' \leq \ell$ or if $\ell+1 \leq i < i' \leq 2k$. This means that there are scalars c and d such that $y = c\gamma + d\gamma'$, where $\gamma = \sum_{j=1}^{\ell} V_{2k}(\boldsymbol{\alpha})^{-1}_{*j}$ and $\gamma' = \sum_{j=\ell+1}^{2k} V_{2k}(\boldsymbol{\alpha})^{-1}_{*j}$. At the same time, $\delta = \sum_1^\ell \overline{\boldsymbol{\vartheta}}_i = \overline{\boldsymbol{\vartheta}}V_{2k}(\boldsymbol{\alpha})\gamma = y \cdot \gamma$ and $-\delta = \sum_{\ell+1}^{2k} \overline{\boldsymbol{\vartheta}}_i = y \cdot \gamma'$. Thus $\|y\|_2^2 = y \cdot (c\gamma + d\gamma') = (c-d)\delta$. Our task is therefore to lower bound $c - d$. We collect our equations for y, δ and $-\delta$ into a 1×2 vector equation:

$$\begin{pmatrix} c & d \end{pmatrix} \begin{pmatrix} \gamma & \gamma' \end{pmatrix}^\dagger \begin{pmatrix} \gamma & \gamma' \end{pmatrix} = \begin{pmatrix} \delta & -\delta \end{pmatrix}$$

Solving for c, d, we get

$$\begin{pmatrix} c & d \end{pmatrix} = \frac{1}{\|\gamma\|_2^2 \cdot \|\gamma'\|_2^2 - (\gamma \cdot \gamma')^2} \begin{pmatrix} \delta & -\delta \end{pmatrix} \begin{pmatrix} \|\gamma'\|_2^2 & -\gamma \cdot \gamma' \\ -\gamma' \cdot \gamma & \|\gamma\|_2^2 \end{pmatrix}.$$

so

$$c - d = \frac{1}{\|\gamma\|_2^2 \cdot \|\gamma'\|_2^2 - (\gamma \cdot \gamma')^2} \begin{pmatrix} \delta & -\delta \end{pmatrix} \begin{pmatrix} \|\gamma'\|_2^2 + \gamma \cdot \gamma' \\ -\gamma' \cdot \gamma - \|\gamma\|_2^2 \end{pmatrix} = \frac{\delta \|\gamma + \gamma'\|_2^2}{\|\gamma\|_2^2 \cdot \|\gamma'\|_2^2 - (\gamma \cdot \gamma')^2}$$

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

Next we examine the denominator. To begin with we upper bound it by $\|\gamma\|_2^2 \cdot \|\gamma'\|_2^2$. (Not much is apt to be lost in this step because by construction γ and γ' are far from identical.) Next we upper bound each of the terms through an interpolation inequality. In what follows we focus on $\|\gamma\|_2^2$, the treatment of $\|\gamma'\|_2^2$ being essentially identical.

The vector γ, interpreted as a polynomial, takes the value 1 on a nonempty set of points $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_\ell$ separated by the positive distance $\boldsymbol{\alpha}_{\ell+1} - \boldsymbol{\alpha}_\ell$ from another nonempty set of points $\boldsymbol{\alpha}_{\ell+1}, \ldots, \boldsymbol{\alpha}_{2k}$ 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. Our inequality (stated in the following lemma) is a converse to this observation. For the purpose of the lemma $2k-1$ is replaced by κ, which may be any (not necessarily odd) positive integer. The proof appears in Appendix A.1

Lemma 2.2. *Let $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{\kappa+1}$, s and ℓ be as above. Let $\gamma(x) = \sum_{i=0}^{\kappa} \gamma_i x^i$ be a real polynomial of degree κ evaluating to 1 at the points $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_\ell$ and evaluating to 0 at the points $\boldsymbol{\alpha}_{\ell+1}, \ldots, \boldsymbol{\alpha}_{\kappa+1}$. Then $\sum_0^\kappa \gamma_i^2 \leq \kappa^2 2^{4\kappa-1} s^{-2\kappa}$.*

We now return to the proof of Theorem 2.1. We have shown that

$$\|y\|_2^2 = (c-d)\delta \geq \frac{\delta^2}{\|\gamma\|_2^2 \cdot \|\gamma'\|_2^2} \geq \frac{\delta^2}{((2k-1)^2 2^{8k-5} s^{-4k+2})^2} = \frac{\delta^2 s^{8k-4}}{(2k-1)^4 2^{16k-10}}.$$

Recall that $\delta \leq 1$ and $\eta \geq \delta s$. So $\|y\|_2^2 \geq \frac{1}{(2k-1)^4 2^{16k-10}} \eta^{8k-4}$. ■

Theorem 1.1 now follows as a simple corollary of the following propositions, whose proofs appear in Appendix A.1. We use V_k, V_{2k}, G, Λ to denote $V_k(\alpha), V_{2k}(\alpha), G(g), \Lambda(\lambda)$ 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. Recall that ζ is the smallest separation between distinct α_i-s.

Proposition 2.3. *If $\|\tilde{g} - g\|_2 \leq \xi$, then $\|G\tilde{\lambda}\|_1 \leq \|G\tilde{\lambda}\|_1 \leq 2^k(k+1)\xi$.*

Proposition 2.4. *For every α_i, $i = 1, \ldots, k$, there exists a $\sigma(i) \in \{1, \ldots, k\}$ such that $\bar{\vartheta}_i |\alpha_i - \tilde{\alpha}_{\sigma(i)}| \leq \frac{8}{\zeta}((k+1)\xi)^{1/k}$.*

Proposition 2.5. *The weights $\bar{\bar{\vartheta}}$ satisfy $\|\bar{\bar{\vartheta}}\tilde{V}_{2k} - \tilde{g}\|_2 \leq \|g - \tilde{g}\|_2 + \frac{8}{\zeta} \cdot (8k)^{3/2}\big((k+1)\xi\big)^{1/k}$.*

So $\|g - \bar{\bar{\vartheta}}\tilde{V}_{2k}\|_2 \leq 2\|g - \tilde{g}\|_2 + \frac{8}{\zeta} \cdot (8k)^{3/2}\big((k+1)\xi\big)^{1/k}$. We can choose ξ small enough using the sample size in Theorem 1.1 so that $\|g - \bar{\bar{\vartheta}}V_{2k}(\tilde{\alpha})\|_2 \leq \frac{1}{(2k-1)^2 2^{8k-5}} \cdot W^{4k-2}$. Coupled with Theorem 2.1, this completes the proof.

Remark: If we are fortunate enough to have a sampling aperture of $\Theta(k^2 \log k)$ instead of $2k-1$, then there is a simple and more straightforward learning algorithm that requires a sample size of only $\Theta(k \log k)$. However, our work is motivated by applications where the sampling aperture is severely constrained. A quadratic increase in the sampling aperture is actually a quadratic loss in the richness of the statistical models that can be learned with the available aperture. (Richness is measured in this section in terms of the number of spikes, or in the next section, for general distributions, in terms of the greatest possible transportation distance between distributions with indistinguishable statistics.)

2.2 Proof of Theorem 1.2: learning arbitrary mixtures

For general mixtures ϑ, the moment map μ' is many-to-one for any finite K. Nevertheless, one can think of its "inverse" as mapping moment vectors to sets of mixtures that produce the same K moments. The main idea of the proof is to bound the Lipschitz constants of the moment map and its inverse.

In this section, we take J to be the interval $(-1, 1)$ for convenience (we exclude coins of type $x \in \pm 1$ for technical reasons; this is not a limitation because in a finite sample it is impossible to distinguish perfectly biased coins from very highly biased ones). Thus, a coin of type x has $\Pr[\text{heads}] = (1-x)/2$.

Recall that for every real r, the transportation norm $\|\cdot\|_{\text{Tran}}$ induces a metric space on the signed measures of trace r on J.

Let $\mathcal{P}$ denote the set of probability measures on J. Let $\mathcal{T}_P(J)$ denote the metric space $(\mathcal{P}, \text{Tran})$: this is a sub-metric space within the trace-1 signed measures. Let $\mathcal{T}_0(J) \subseteq \bar{\mathcal{T}}_0(J)$ consist of the trace-0 bounded signed measures on J which assign measure ≤ 1 to any set. Observe that $\mathcal{T}_0(J) = \mathcal{T}_P(J) - \mathcal{T}_P(J) = (\mathcal{P} - \mathcal{P}, \|\cdot\|_{\text{Tran}})$, where $\mathcal{P} - \mathcal{P}$ is Minkowsky sum.

Observation 2.6. *Let S be the definite integration operator on J defined by $(S(\vartheta))(x) = \vartheta((-1,x))$, and let $\vartheta \in \bar{\mathcal{T}}_0(J)$. Then, $\|\vartheta\|_{\mathrm{Tran}} = \|S(\vartheta)\|_1$.*

Recall our notation of the empirical statistics $\tilde{f} = \left(\widetilde{\mathsf{freq}_0}, \frac{\widetilde{\mathsf{freq}_1}}{K}, \frac{\widetilde{\mathsf{freq}_2}}{\binom{K}{2}}, \ldots, \widetilde{\mathsf{freq}_K}\right)$ and the moment vector $f = \mu(\vartheta) = \left(\mathsf{freq}_0, \frac{\mathsf{freq}_1}{K}, \frac{\mathsf{freq}_2}{\binom{K}{2}}, \ldots, \mathsf{freq}_K\right)$. The linear mapping $\mu : \mathcal{T}_P(J) \to \ell_2^{K+1}$ is defined by the maps $\mathsf{freq}_i(\vartheta) = \int_{-1}^{1} \binom{K}{i} P_i(x) \cdot d\vartheta(x)$ for each $0 \le i \le K$, with $P_i(.)$ being the polynomial $P_i(x) = \left(\frac{1-x}{2}\right)^i \left(\frac{1+x}{2}\right)^{K-i}$, In short, $f_i = \mu(\vartheta)_i = 2^{-K} \int_{-1}^{1} (1-x)^i (1+x)^{K-i}\, d\vartheta(x)$.

By linearity, μ extends to a mapping $\mu : \mathcal{T}_0(J) \to \ell_2^{K+1}$ (setting $\mu(\vartheta - \vartheta') = \mu(\vartheta) - \mu(\vartheta')$). By Observation 2.6, the linear map $S : \bar{\mathcal{T}}_0(J) \to L_1(J)$ is an isometry. Consider the restriction of S to $\mathcal{T}_0(J)$; define $\mathbb{J}_0 \subset L_1(J)$ to be the image of this restriction, and let $D : \mathbb{J}_0 \to \mathcal{T}_0(J)$ denote the inverse of S on $\mathbb{J}_0$. Note that $\mathbb{J}_0$ consists of the functions $g \in L_1(J)$ of total variation at most 2, for which $g(-1) = g(1) = 0$.

We first show that the observed moments suffice to determine a close approximation to ϑ. The mapping $\mu \circ D : \mathbb{J}_0 \to \ell_2^{K+1}$ is far from isometric, being many-to-one. It nonetheless has very nice metric properties. These are established in Lemmas 2.7 and 2.8, which yield Theorem 2.9 as a simple corollary, and in Proposition 2.10, which is used for the algorithmic reconstruction. Lemma 2.7 establishes that the preimage of any point under $\mu \circ D$ has small diameter. Lemma 2.8 establishes that the preimage changes only gradually as the point is moved.

Lemma 2.7. $\sup\{\|g\|_1 : g \in (\mu \circ D)^{-1}(\vec{0})\} \le \sqrt{\frac{8}{\pi(K+1)}}$.

For a metric space (M, d) the induced Hausdorff metric on subsets, which we denote $(2^M, d)$, is given by $d(S,T) = \sup_{s_1 \in S, t_1 \in T} \inf_{s_2 \in S, t_2 \in T} \max\{d(s_1,t_2), d(t_1,s_2)\}$. Consider $(\mu \circ D)^{-1}$ as a mapping from ℓ_2^{K+1} to the Hausdorff metric on subsets of $L_1(J)$.

Lemma 2.8. $\|(\mu \circ D)^{-1}\|_{\mathrm{Lip}} \le C$ *for some finite* $C = C(K) = K^{O(K)}$.

Theorem 2.9. *If $\|\tilde{f} - f\|_2 \le \delta$ then any reconstructed measure $\tilde{\vartheta}$ such that $\mu\tilde{\vartheta} = \tilde{f}$ satisfies $\|\tilde{\vartheta} - \vartheta\|_{\mathrm{Tran}} \le \delta C(K) + \sqrt{8/(\pi(K+1))}$.*

Proof Sketch of Lemma 2.7. The proof is fairly technical but the approach to it is this. Consider $g \in (\mu \circ D)^{-1}(\vec{0})$. Let $T_n(x)$ denote the Chebyshev polynomials of the first kind and let $U_n(x) = \frac{1}{n+1} \frac{dT_{n+1}(x)}{dx}$ denote the Chebyshev polynomials of the second kind. We will expand $g \in \mathbb{J}_0$ in the basis of the Chebyshev polynomials $\{U_n\}$ and obtain a bound on these Chebyshev coefficients (by a method inspired by a classical bound on the Fourier coefficients of functions of bounded variation). The proof is in Appendix A.2. ∎

Proof of Lemma 2.8. Let $F = \mu(\mathcal{T}_P(J))$ be the image of $\mathcal{T}_P(J)$ under the linear transformation μ. We wish to show that there exists a finite $C(K)$ such that given any $f_1, f_2 \in F$ and $g_1 \in \mathbb{J}_P$ that satisfies $f_1 = (\mu \circ D)(g_1)$, there exists $g_2 \in \mathbb{J}_P$ such that $f_2 = (\mu \circ D)(g_2)$ and $\|g_1 - g_2\|_1 \le C(K) \cdot \|f_1 - f_2\|_2$. Since $\mu \circ D$ is linear, it suffices to show the equivalent claim that for every $f \in \mu(\mathcal{T}_0(J))$ there exists $g \in \mathbb{J}_0$ such that $f = (\mu \circ D)(g)$ and $\|g\|_1 \le C(K) \cdot \|f\|_2$.

Let $H_0(x), \ldots, H_K(x)$ denote a basis for the space of polynomials of degree at most K, that is dual on J to the polynomials $P_0(x), P_1(x), \ldots, P_K(x)$. In other words, $\int_{-1}^{1} H_i(x) \cdot P_j(x)\ dx = \delta_{ij}$. Set $g = S\left(\sum_{i=0}^{K} f_i H_i\right)$. Then $(\mu \circ D)(g) = \mu\left(\sum_{i=0}^{K} f_i H_i\right) = \sum_{i=0}^{K} f_i \mu(H_i) = f$. Let L_n denote the Legendre polynomials, scaled so that they are orthonormal on J. Let B denote the $(K+1) \times (K+1)$ change-of-basis matrix from $S(H_0), \ldots, S(H_K)$ to $L_1, \ldots, L_{K+1}$; B is defined by $B_{in} = \int_{-1}^{1} L_n(y) \cdot S(H_i)(y)\ dy$. (Note $\sum_i \gamma_i S(H_i) = \sum_n L_n \int_{-1}^{1} L_n(x) \sum_i \gamma_i S(H_i)(x)\ dx = \sum_n L_n \sum_i \gamma_i B_{in}$.) Notice that $g(x)$ is a constant-free

polynomial of degree at most $K+1$, therefore $g \in \mathrm{Span}\{L_1, \ldots, L_{K+1}\}$. Thus,

$$\begin{aligned} g(x) &= \sum_{n=1}^{K+1} L_n(x) \cdot \int_{-1}^{1} L_n(y) \cdot g(y)\, dy = \sum_{n=1}^{K+1} L_n(x) \cdot \int_{-1}^{1} L_n(y) \cdot S\left(\sum_{i=0}^{K} f_i H_i(y)\right) dy \\ &= \sum_{n=1}^{K+1} L_n(x) \cdot \sum_{i=0}^{K} f_i \int_{-1}^{1} L_n(y) \cdot S(H_i)(y)\, dy = \sum_{n=1}^{K+1} L_n(x) \cdot \sum_{i=0}^{K} f_i B_{in}. \end{aligned}$$

Therefore, $\|g\|_2^2 = \|fB\|_2^2$, so $\|g\|_1 \le \sqrt{2} \cdot \|g\|_2 \le \sqrt{2} \cdot \|f\|_2 \cdot \|B\|_{\mathrm{op}}$, where $\|\cdot\|_{\mathrm{op}}$ is the matrix operator norm. Finally, take $C(K) = \sqrt{2} \cdot \|B\|_{\mathrm{op}}$. (In this abstract we omit the argument that $\|B\|_{\mathrm{op}} \in K^{O(K)}$.) ■

We now complete the proof of Theorem 1.2 by giving an algorithm to invert the noisy moment map. The key to our algorithm is an upper bound on the Lipschitz constant of $\mu \circ D$.

Proposition 2.10. $\|\mu \circ D\|_{\mathrm{Lip}} \le K$. *(The proof appears in Appendix A.2.)*

This proposition, in conjunction with Theorem 2.9, allows us to restrict our search for ϑ to any subset of $\mathcal{T}_P(J)$ dense enough to constitute a good "covering code." More specifically, let $\varepsilon > 0$, and let $S \subseteq \mathcal{T}_P(J)$ be such that $\sup_{\vartheta_1 \in \mathcal{T}_P(J)} \inf_{\vartheta_2 \in S} \|\vartheta_1 - \vartheta_2\|_{\mathrm{Tran}} \le \varepsilon/(2K \cdot C(K))$. Let the sample size be large enough that with high probability $\|\tilde{f} - f\|_2 \le \varepsilon/(2C(K))$. The covering code condition and the lemma ensure that there exists a $\vartheta' \in S$ such that $\|\mu(\vartheta') - \tilde{f}\|_2 \le \varepsilon/(2C(K))$. So (if the high probability event occurs), ϑ' is such that $\|\mu(\vartheta') - f\|_2 \le \varepsilon/C(K)$. Then by Theorem 2.9, $\|\vartheta' - \vartheta\|_{\mathrm{Tran}} \le \varepsilon + \sqrt{8/(\pi(K+1))}$.

It remains to choose a suitable covering code and to provide an algorithm to search it for a measure ϑ' satisfying $\|\mu(\vartheta') - \tilde{f}\|_2 \le \varepsilon/(2C(K))$. There is a fairly obvious (and far from unique) covering code: pick $\varepsilon = \sqrt{8/(\pi(K+1))}$, set $s = \lceil K \cdot C(K)/\varepsilon \rceil$, and specify as the covering code the collection of measures supported on the finite set of coin types $\{x_j\}_{-s \le \text{integer } j \le s-1}$, where $x_j = \frac{j+1/2}{s}$. Let A denote the $2s \times (K+1)$ matrix $A_{ji} = P_i(x_j)$. To find a suitable measure ϑ', solve the convex program

$$\text{minimize } \Big\{ \|\vartheta' A - \tilde{f}\|_2 : \ \ \vartheta' \ge 0 \Big\}.$$

This can be done in time exponential in K (i.e., polynomial in the sample size). Thus we have demonstrated Theorem 1.2.

3 Proof of Theorem 1.3

Our reduction uses spectral methods. In particular, we use results on the concentration of the eigenvalues of random matrices [24, 2, 41]. Previous empirical and theoretical results on spectral analysis of data [7, 18, 37, 33, 8, 30, 9, 20, 28, 11, 34, 5, 39, 1, 27, 14] are somewhat related to our work, though our algorithm differs from previous uses of these tools. We note that a more complicated algorithm gives slightly tighter bounds on the required sample size; this is deferred to the full version of the paper.

Let $\bar{p}, \bar{q} \in \Delta_n$ and consider a mixture $\boldsymbol{\theta}$ of distributions of the form $a\bar{p} + (1-a)\bar{q}$. Let $r \in \Delta_n$ denote the expected face probabilities with respect to $\boldsymbol{\theta}$, and let σ^2 denote the linear variance of $\boldsymbol{\theta}$ (along the supporting interval $[\bar{p}, \bar{q}]$). Put $v = \frac{\sigma}{2} \cdot (\bar{q} - \bar{p})$. Note that for all $s \in [n]$, $|v_s| \le r_s$. Let M denote the pairwise correlation matrix for the underlying distribution on pairs of rolls. I.e., for $s, t \in [n]$, M_{st} is half the probability that two rolls of a die produce s, t in any order.

Proposition 3.1. $M = \frac{1}{2}\left((r+v)(r+v)^T + (r-v)(r-v)^T\right)$.

We will use the notation $p = r - v$ and $q = r + v$. So $r = \frac{1}{2}(p+q)$ and $v = \frac{1}{2}(q-p)$. Let $\zeta = \frac{1}{2}\|p-q\|_1 = \sum_{s=1}^{n} |v_s|$ be the total variation distance between p and q. Notice that for every $s \in [n]$, $r_s = \sum_{t=1}^{n} M_{st}$ (because $\sum_{t=1}^{n} v_t = 0$). Thus, if we have M, we can compute trivially r and also v, which is the principal eigenvector of $M - rr^T = vv^T$. Sampling m dice, two rolls per die, we get an estimate $\tilde{M}$ for M. I.e., $\tilde{M}$ is an $n \times n$ symmetric matrix, where $\tilde{M}_{st}$ is half the fraction of die roll pairs that fell on s and t (in any order). Using $\tilde{M}$ (a very noisy estimator of M), our reduction algorithm computes vectors $\tilde{r}$ and $\tilde{v}$, and outputs the endpoints $\tilde{p}, \tilde{q}$ of the interval $\tilde{r} + \mathrm{Span}(\tilde{v}) \cap \Delta_n$. We show that the line $\tilde{r} + \mathrm{Span}(\tilde{v})$ is sufficiently close to the line $r + \mathrm{Span}(v)$ to enable the computation of a good estimate for $\boldsymbol{\theta}$. We now describe in detail how to compute $\tilde{r}$ and $\tilde{v}$. Let $c > 0$ be a sufficiently large constant. Fix $\delta, \epsilon, S > 0$ such that $\delta \le \frac{\zeta}{c}$, $\epsilon \le \frac{\delta}{\log(n/\delta)}$, $S \ge \frac{c}{\epsilon^4}$, and $m \ge \frac{c}{\delta\epsilon^8} \cdot n \log n + cnS^2$ (recall m is the sample size). Our reduction proceeds as follows.

(1) For every $s \in [n]$, compute $\tilde{r}_s = \sum_{t=1}^{n} \tilde{M}_{st}$.

(2) For $j = 1, 2, \ldots$, compute $I_j = \{s : \ 2^{-j} \le \tilde{r}_s < 2^{-j+1}\}$. (We assume that for all $s \in [n]$, $\tilde{r}_s < 1$, otherwise the problem is trivial.)

For ease of analysis, we now consider another empirical version $\hat{M}$ of M that is independent of $\tilde{M}$. There are $\hat{m}$ die samples from $\boldsymbol{\theta}$, where $\hat{m}$ is distributed Poisson (this can be emulated from the real data). Each die is rolled twice. We put $\hat{M}_{st}$ to be half the fraction of die roll pairs that landed on s and t in any order.

(3) For every $I = I_j$, where $j = 1, 2, \ldots$, compute $\hat{V}_{I\times I} = \hat{M}_{I\times I} - \tilde{r}_I \tilde{r}_I^T$.

(4) For every $I = I_j$, $j = 1, 2, \ldots$, compute $\hat{v}_I \in \mathbb{R}^I$, the principal eigenvector of $\hat{V}_{I\times I}$, with $\|\hat{v}_I\|_2 = 1$, and compute $\lambda(I) = \lambda_1(\hat{V}_{I\times I})$.

(5) Compute $J = \left\{ j \in \mathbb{N} : \ \|\tilde{r}_{I_j}\|_1 \ge \epsilon \quad \wedge \quad \lambda(I_j) \ge \frac{\epsilon^2}{2\cdot|I_j|} \right\}$ (we show that $J \neq \emptyset$ in Lemma B.6).

(6) For any $I \subseteq [n]$ and a vector $x \in \mathbb{R}^I$, put $T^+(x) = \{s \in I : \ x_s > 0\}$. For $j \in \mathbb{N}$, we use the notation $T_j^+ = T^+(v_{I_j})$, $T_j^- = T^+(-v_{I_j})$, $\hat{T}_j^+ = T^+(\hat{v}_{I_j})$, and $\hat{T}_j^- = T^+(-\hat{v}_{I_j})$. Pick an arbitrary $j_0 \in J$. Without loss of generality, assume that $v_{I_{j_0}}^T \hat{v}_{I_{j_0}} > 0$ and $\|\hat{v}_{\hat{T}_{j_0}^+}\|_2^2 \ge \frac{1}{2}\|\hat{v}_{I_{j_0}}\|_2^2 = \frac{1}{2}$. For every $j \in J$ put

$$e_j = \begin{cases} \sum_{s\in\hat{T}_j^+} \sum_{t\in\hat{T}_{j_0}^+} (\tilde{M}_{st} - \tilde{r}_s\tilde{r}_t) & \text{if } \|\hat{v}_{\hat{T}_j^+}\|_2^2 \ge \frac{1}{2}; \\ \sum_{s\in\hat{T}_j^-} \sum_{t\in\hat{T}_{j_0}^+} (\tilde{r}_s\tilde{r}_t - \tilde{M}_{st}) & \text{otherwise.} \end{cases}$$

(7) Finally, compute $\tilde{v} \in \mathbb{R}^n$ as follows:

$$\tilde{v}_s = \begin{cases} \sqrt{\lambda(I_j)} \cdot \hat{v}_s & \text{if } \exists j \in J \text{ s.t. } s \in I_j \text{ and } e_j > 0; \\ -\sqrt{\lambda(I_j)} \cdot \hat{v}_s & \text{if } \exists j \in J \text{ s.t. } s \in I_j \text{ and } e_j < 0; \\ 0 & \text{otherwise.} \end{cases}$$

Let $T^+ = T^+(v)$, $T^- = T^+(-v)$, $\tilde{T}^+ = T^+(\tilde{v})$, and $\tilde{T}^- = T^+(-\tilde{v})$. Theorem 1.3 is an immediate corollary of Theorem 3.2: φ induces a low-cost transport between measures on $[p, q]$ and measures on $[\tilde{p}, \tilde{q}]$, and $\psi_{\tilde{p},\tilde{q}}$ maps measures on $[\tilde{p}, \tilde{q}]$ isometrically to $[0, 1]$. The proof appears in Appendix B.

Theorem 3.2. *There exists a constant $\kappa > 0$ and a mapping $\varphi : \mathbb{R} \to \mathbb{R}$ that satisfy the following conditions with high probability. For every $\alpha \in \mathbb{R}$ such that $r + \alpha \cdot v$ is a point in the simplex,*

$$|\alpha - \varphi(\alpha)| \le \zeta^{-2} \cdot \kappa \cdot \log(n/\delta) \cdot \sqrt{1/S + \sqrt{(n \ln n)/(\delta \cdot m)}}; \tag{1}$$

$$\frac{\sum_{s\in\tilde{T}^+}(r_s+\alpha\cdot v_s)}{\sum_{s\in\tilde{T}^+\cup\tilde{T}^-}(r_s+\alpha\cdot v_s)}=\frac{\sum_{s\in\tilde{T}^+}(\tilde{r}_s+\varphi(\alpha)\cdot\tilde{v}_s)}{\sum_{s\in\tilde{T}^+\cup\tilde{T}^-}(\tilde{r}_s+\varphi(\alpha)\cdot\tilde{v}_s)}; \tag{2}$$

$$\|(r+\alpha\cdot v)-(\tilde{r}+\varphi(\alpha)\cdot\tilde{v})\|_1\leq\kappa\cdot\left(\zeta^{-1}\cdot\delta+\zeta^{-2}\cdot\log(n/\delta)\cdot\sqrt{1/S+\sqrt{(n\ln n)/(\delta\cdot m)}}\right). \tag{3}$$

References

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

[2] N. Alon, M. Krivelevich, and V. H. Vu. Concentration of eigenvalue of random matrices. *Israel Math. Journal*, 131:259–267, 2002.

[3] N. Alon and J. H. Spencer. *The probabilistic method.* Wiley, 2nd edition, 2000.

[4] S. Arora and R. Kannan. Learning mixtures of separated nonspherical gaussians. *Ann. Appl. Probab.*, 15(1A):69–92, 2005. Preliminary version appeared in *Proceedings of 33rd STOC*, 2001.

[5] Y. Azar, A. Fiat, A. R. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proc. STOC '01*, pages 619–626, 2001.

[6] T. Batu, S. Guha, and S. Kannan. Inferring mixtures of markov chains. In *Proc. 17th Ann. Conf. on Computational Learning Theory*, pages 186–199, 2004.

[7] R. Bopanna. Eigenvalues and graph bisection: an average-case analysis. In *Proc. 28th IEEE Symp. on Foundations of Computer Science*, pages 280–285, 1987.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[9] S. Chakrabarti, B. E. Dom, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the web's link structure. *Computer*, 32(8):60–67, 1999.

[10] K. Chaudhuri, E. Halperin, S. Rao, and S. Zhou. A rigorous analysis of population stratification with limited data. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, January 2007.

[11] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

[12] R. Cottle, E. Johnson, and R. Wets. George B. Dantzig (1914-2005). *Notices of the AMS*, 54(3):344–362, March 2007.

[13] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002.

[14] A. Dasgupta, J. Hopcroft, R. Kannan, and P. Mitra. Spectral clustering with limited independence. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, January 2007.

[15] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler. On learning mixtures of heavy-tailed distributions. In *Proc. 46th Ann. IEEE Symp. on Foundations of Computer Science*, pages 491–500, 2005.

[16] S. Dasgupta. Learning mixtures of gaussians. In *Proc. IEEE Symp. on Foundations of Computer Science*, pages 634–644, 1999.

[17] S. Dasgupta and L.J. Schulman. A two-round variant of em for gaussian mixtures. In *Conference in Uncertainty in Artificial Intelligence*, pages 143–151, 2000.

[18] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.

[19] J. L. Doob. *Measure theory*. Springer-Verlag, 1994.

[20] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proceedings of the 10th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 1999.

[21] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th Symp. Found. Comp. Sci.*, 2005.

[22] J. Feldman, R. O'Donnell, and R. Servedio. Pac learning mixtures of axis-aligned gaussians with no separation assumption. In *Proc. 19th Ann. Conf. on Computational Learning Theory*, pages 20–34, 2006.

[23] Y. Freund and M. Mansour. Estimating a mixture of two product distributions. In *Proc. 12th Ann. Conf. Computational Learning Theory*, pages 183–192, 1999.

[24] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.

[25] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. Birkhauser, 1999.

[26] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *Proc. Int'l Joint Conf. in Artificial Intelligence*, 1999.

[27] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proc. COLT*, pages 444–457, 2005.

[28] R. Kannan, S. Vempala, and A. Vetta. On clusterings: good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.

[29] 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.

[30] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. of the ACM*, 46, 1999.

[31] J. Kleinberg and M. Sandler. Using mixture models for collaborative filtering. In *Proc. 36th STOC*, 2004.

[32] J. Kleinberg and M. Sandler. Using mixture models for collaborative filtering. *J. Computer and System Sciences*, 74:49–69, 2008.

[33] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Recommendation systems: A probabilistic analysis. *J. Comput. Syst. Sci.*, 63(1):42–61, 2001.

[34] F. McSherry. Spectral partitioning of random graphs. In *Proc. IEEE Symp. on Foundations of Computer Science*, pages 529–537, 2001.

[35] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden markov models. In *Proc. 37th Ann. ACM Symp. on Theory of Computing*, 2005.

[36] V. Y. Pan. Optimal and nearly optimal algorithms for approximating polynomial zeros. *Computers & Mathematics with Applications*, 31(12):97–138, 1996. Preliminary version appeared in *Proceedings of 27th STOC*, 1995.

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

[38] M. Taibleson. Fourier coefficients of functions of bounded variation. *Proceedings of the American Mathematical Society*, 18(4):766, August 1967.

[39] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. *J. Computer and System Sciences*, 68(4):841–860, 2004. Originally in 43rd IEEE FOCS 2002.

[40] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[41] V. H. Vu. Spectral norm of random matrices. In *Proc. 37th Ann. ACM Symp. on Theory of Computing*, 2005.

A Proofs from Section 2

A.1 Proofs from Section 2.1

Proof of Lemma 2.2. 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:

$$\gamma(x) = \prod_{j=2}^{\kappa+1} \frac{x - \boldsymbol{\alpha}_j}{\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_j}$$

For $0 \le i \le \kappa$ let $e_i^\kappa(\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{\kappa+1})$ be the i'th *elementary symmetric mean*,

$$e_i^\kappa(\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{\kappa+1}) = \frac{1}{\binom{\kappa}{i}} \sum_{S \in \binom{\{2,\ldots,\kappa+1\}}{i}} \prod_{j \in S} \boldsymbol{\alpha}_j$$

and observe that for all i, $0 \le e_i^\kappa(\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{\kappa+1}) \le 1$. Now

$$\gamma(x) = \left(\prod_{j=2}^{\kappa+1} \frac{1}{\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_j} \right) \sum_{i=0}^{\kappa} (-1)^{\kappa-i} \binom{\kappa}{i} e_{\kappa-i}^\kappa(\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{\kappa+1}) x^i$$

So $\sum \gamma_i^2 = \left(\prod_{j=2}^{\kappa+1} \frac{1}{\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_j} \right)^2 \sum \left(\binom{\kappa}{i} e_i^\kappa(\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{\kappa+1}) \right)^2 \le s^{-2\kappa} \sum \binom{\kappa}{i}^2 = \binom{2\kappa}{\kappa} s^{-2\kappa}$.

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 $\sum \gamma_i^2 \le 4\binom{2\kappa}{\kappa} s^{-2\kappa}$.

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 $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_\ell$ or among $\boldsymbol{\alpha}_{\ell+1}, \ldots, \boldsymbol{\alpha}_{\kappa+1}$ 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 $\boldsymbol{\alpha}'_1 \le \ldots \le \boldsymbol{\alpha}'_{\kappa-1}$ denoting the roots of $\partial\gamma/\partial x$, that for $j < \ell$, $\boldsymbol{\alpha}_j \le \boldsymbol{\alpha}'_j \le \boldsymbol{\alpha}_{j+1}$, and for $j \ge \ell$, $\boldsymbol{\alpha}_{j+1} \le \boldsymbol{\alpha}'_j \le \boldsymbol{\alpha}_{j+2}$. In particular, none of the roots fall in the interval $(\boldsymbol{\alpha}_\ell, \boldsymbol{\alpha}_{\ell+1})$. For some constant C we can write $\partial\gamma/\partial x = C \prod_{j=0}^{\kappa-1} (x - \boldsymbol{\alpha}'_j)$ (with $\operatorname{sign}(C) = (-1)^{1+\kappa-\ell}$). Observe that $\int_{\boldsymbol{\alpha}_\ell}^{\boldsymbol{\alpha}_{\ell+1}} \frac{\partial\gamma}{\partial x}(x)\, dx = -1$. So $(-1)^{1+\kappa-\ell}/C = \int_{\boldsymbol{\alpha}_\ell}^{\boldsymbol{\alpha}_{\ell+1}} (-1)^{\kappa-\ell} \prod_{j=0}^{\kappa-1} (x - \boldsymbol{\alpha}'_j)\, dx$. Observe that if for any $j < \ell$, $\boldsymbol{\alpha}'_j$ is increased, or if for any $j \ge \ell$, $\boldsymbol{\alpha}'_j$ is decreased, then the integral decreases. So $(-1)^{1+\kappa-\ell}/C \ge \int_{\boldsymbol{\alpha}_\ell}^{\boldsymbol{\alpha}_{\ell+1}} (-1)^{\kappa-\ell} (x - \boldsymbol{\alpha}_\ell)^{\ell-1} (x - \boldsymbol{\alpha}_{\ell+1})^{\kappa-\ell}\, dx$. This is a definite integral that can be evaluated in closed form.

$$\int_{\boldsymbol{\alpha}_\ell}^{\boldsymbol{\alpha}_{\ell+1}} (-1)^{\kappa-\ell} (x - \boldsymbol{\alpha}_\ell)^{\ell-1} (x - \boldsymbol{\alpha}_{\ell+1})^{\kappa-\ell}\, dx = (\boldsymbol{\alpha}_{\ell+1} - \boldsymbol{\alpha}_\ell)^\kappa (\ell-1)!(\kappa-\ell)!/\kappa!$$

$$(-1)^{1+\kappa-\ell} C \le \frac{\kappa!}{s^\kappa (\ell-1)!(\kappa-\ell)!}$$

The sum of squares of coefficients of $\frac{\partial\gamma}{\partial x}$ is $C^2 \sum_{i=0}^{\kappa-1} \binom{\kappa-1}{i}^2 (e_i^{\kappa-1}(\boldsymbol{\alpha}'_1, \ldots, \boldsymbol{\alpha}'_{\kappa-1}))^2 \le C^2 \binom{2\kappa-2}{\kappa-1}$. 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 $C \sum_{i=0}^{\kappa-1} \binom{\kappa-1}{i} = C 2^\kappa$. So $|\gamma_0| \le C 2^\kappa$. Consequently,

$$\begin{aligned}
\sum_0^\kappa \gamma_i^2 \le C^2 \left(\binom{2\kappa-2}{\kappa-1} + 2^{2\kappa} \right) &\le \left(\binom{2\kappa-2}{\kappa-1} + 2^{2\kappa} \right) \left(\frac{\kappa!}{(\ell-1)!(\kappa-\ell)!} \right)^2 s^{-2\kappa} \\
&\le 5\kappa^2 2^{2\kappa-2} \binom{\kappa-1}{\ell-1}^2 s^{-2\kappa} \\
&\le 5\kappa^2 2^{4\kappa-4} s^{-2\kappa},
\end{aligned}$$

which completes the proof of the lemma. ■

Proof of Proposition 2.3. 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 latter inequality follows since $P_\lambda = \prod_{i=1}^{k}(x - \alpha_i)$; hence, $|\lambda_i| \leq \binom{k}{i}$. So for any $i = 1, \ldots, k$, $\left|(G - \tilde{G})_i \cdot \lambda\right| \leq \|\lambda\|_2 \|G_i - \tilde{G}_i\|_2 \leq 2^k\xi$. Thus, λ is a feasible solution to (P), which implies that $\|\tilde{\lambda}\|_1 \leq 2^k$. We have $\|G\tilde{\lambda}\|_1 \leq \|\tilde{G}\tilde{\lambda}\|_1 + \|(G - \tilde{G})\tilde{\lambda}\|_1 \leq 2^k\xi + \|(G - \tilde{G})\tilde{\lambda}\|_1$. For any $i = 1, \ldots, k$, $\left|(G - \tilde{G})_i \cdot \tilde{\lambda}\right| \leq \|G_i - \tilde{G}_i\|_2 \|\tilde{\lambda}\|_2 \leq 2^k\xi$, so $\|G\tilde{\lambda}\|_1 \leq 2^k(k+1)\xi$. ■

Proof of Proposition 2.4. Since $\|G\tilde{\lambda}\|_2 \leq 2^k(k+1)\xi$ (by Proposition 2.3), we have equivalently that the $\|.\|_2$ norm of $g\tilde{\Lambda} = \bar{\vartheta}V_{2k}\tilde{\Lambda}$ is at most $2^k(k+1)\xi$. We may write $\bar{\vartheta}V_{2k}\tilde{\Lambda}$ as

$$\vartheta V_{2k}\tilde{\Lambda} = \begin{pmatrix} \bar{\vartheta}_1 & \cdots & \bar{\vartheta}_k \end{pmatrix} \begin{pmatrix} P_{\tilde{\lambda}}(\alpha_1) & \alpha_1 P_{\tilde{\lambda}}(\alpha_1) & \cdots & \alpha_1^{k-1} P_{\tilde{\lambda}}(\alpha_1) \\ P_{\tilde{\lambda}}(\alpha_2) & \alpha_2 P_{\tilde{\lambda}}(\alpha_2) & \cdots & \alpha_2^{k-1} P_{\tilde{\lambda}}(\alpha_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_{\tilde{\lambda}}(\alpha_k) & \alpha_k P_{\tilde{\lambda}}(\alpha_k) & \cdots & \alpha_k^{k-1} P_{\tilde{\lambda}}(\alpha_k) \end{pmatrix}$$

which is equal to $\vartheta' V_k(\alpha)$ where $\vartheta' = \big(\bar{\vartheta}_1 P_{\tilde{\lambda}}(\alpha_1), \cdots, \bar{\vartheta}_k P_{\tilde{\lambda}}(\alpha_k)\big)$. Thus, we are given that $\|\vartheta' V_k\|_2 \leq 2^k(k+1)\xi$. Proceeding as in the proof of Theorem 2.1, we can also obtain the lower bound

$$\|\vartheta' V_k\|_2 \;\geq\; \max_i\Big(|\vartheta'_i| \cdot \prod_{j \neq i} \frac{|\alpha_i - \alpha_j|}{1 + \alpha_j}\Big) \;\geq\; \max_i\Big(\bar{\vartheta}_i\Big(\frac{\zeta}{2}\Big)^{k-1} \prod_{j=0}^{k} |\alpha_i - \bar{\alpha}_j|\Big) \;\geq \max_i\Big(\bar{\vartheta}_i\Big(\frac{\zeta}{2}\Big)^{k-1} \prod_{j=0}^{k} |\alpha_i - \mathrm{Re}(\bar{\alpha}_j)|\Big).$$

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,

$$\prod_{j=0}^{k} |\alpha_i - \bar{\alpha}_j| = \big((\alpha_i - a)^2 + b^2\big) \cdot \prod_{j=0,\, j\neq \ell, \ell'}^{k} |\alpha_i - \bar{\alpha}_j| \geq (\alpha_i - a)^2 \cdot \prod_{j=0\, j\neq \ell, \ell'}^{k} |\alpha_i - \bar{\alpha}_j|.$$

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

$$\|\vartheta' V_k\|_2 \;\geq\; \max_i\Big(\bar{\vartheta}_i\Big(\frac{\zeta}{2}\Big)^{k-1} \prod_{j=0}^{k} \big||\alpha_i - \tilde{\alpha}_j| - \epsilon\big|\Big).$$

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

Proof of Proposition 2.5. Let $\eta = \frac{8}{\zeta} \cdot \big((k+1)\xi\big)^{1/k}$. We will bound $\|\overline{\tilde{\vartheta}}\tilde{V}_{2k} - \tilde{g}\|_2$ by exhibiting a solution $y \in [0,1]^k$, $\|y\|_1 = 1$ such that $\|y\tilde{V}_{2k} - \tilde{g}\|_2 \leq \|g - \tilde{g}\| + (8k)^{3/2}\eta$. Let σ be the function whose existence is proved in Lemma 2.4. 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 $g - y\tilde{V}_{2k} = \bar{\vartheta}V_{2k} - y\tilde{V}_{2k} = \sum_{i=1}^{k} \vartheta_i (V_{2k,i} - \tilde{V}_{2k,\sigma(i)})$ where $V_{2k,i}$ and $\tilde{V}_{2k,i}$ denote respectively the i-th rows of V_{2k} and $\tilde{V}_{2k}$. 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, $\|g - y\tilde{V}_{2k}\|_2 \leq (8k)^{3/2}\eta$. ■

A.2 Proofs from Section 2.2

Proof of Proposition 2.10. Since D is an isometry, the lemma is equivalent to showing that μ has Lipschitz constant bounded by K on $\mathcal{T}_0(J)$. To establish this, we bound how far $\mu(\vartheta)$ can move (in ℓ_2) when ϑ is changed by transporting a delta function. Specifically, for $a \geq 0, -1 < z_0 \leq z_1 < 1$, and with δ_z denoting the (Dirac) measure assigning measure 1 to sets containing z and 0 to other sets, write $\vartheta(a, z_0, z_1) = \vartheta + a(\delta_{z_1} - \delta_{z_0})$. Now we show $\left\|\mu(\vartheta) - \mu(\vartheta + a(\delta_{z_1} - \delta_{z_0}))\right\|_2 \leq Ka(z_1 - z_0)$, which follows from the following claim:

$$\left\| \frac{\partial}{\partial u}\Big|_{u=0} (\mu(\vartheta) - \mu(\vartheta + a(\delta_{z+u} - \delta_z))) \right\|_2 \leq Ka$$

We establish this as follows:

$$\begin{aligned}
LHS &= \sqrt{\sum_i \left(\frac{\partial}{\partial u}\Big|_{u=0} \left(\int_{-1}^{1} \left(\frac{1-x}{2}\right)^i \left(\frac{1+x}{2}\right)^{K-i} d(\vartheta + a(\delta_{z+u} - \delta_z))(x) \right) \right)^2} \\
&= \sqrt{\sum_i \left(a \frac{\partial}{\partial u}\Big|_{u=0} \left(\left(\frac{1-z-u}{2}\right)^i \left(\frac{1+z+u}{2}\right)^{K-i} \right) \right)^2} \\
&= a\sqrt{\sum_i \left(\left(\frac{1-z}{2}\right)^i \left(\frac{1+z}{2}\right)^{K-i} \left(\frac{2(K-i)}{1+z} - \frac{2i}{1-z} \right) \right)^2} \\
&= a\sqrt{4\sum_i \left(\frac{1-z}{2}\right)^{2i} \left(\frac{1+z}{2}\right)^{2(K-i)} \left(\frac{K-i}{1+z} - \frac{i}{1-z} \right)^2} \\
&= a\sqrt{4\sum_i \left(\frac{1-z}{2}\right)^{2i} \left(\frac{1+z}{2}\right)^{2(K-i)} \left(\frac{K-2i-zK}{(1+z)(1-z)} \right)^2} \\
&\leq a\sqrt{4\sum_i \left(\frac{1-z}{2}\right)^{2i} \left(\frac{1+z}{2}\right)^{2(K-i)} \left(\frac{2K}{(1+z)(1-z)} \right)^2} \\
&= Ka\sqrt{\sum_i \left(\frac{1-z}{2}\right)^{2i-2} \left(\frac{1+z}{2}\right)^{2(K-i-1)}} \\
&= Ka\sqrt{\frac{\left(\frac{1-z}{2}\right)^{2K} - \left(\frac{1+z}{2}\right)^{2K}}{\left(\frac{1-z}{2}\right)^2 - \left(\frac{1+z}{2}\right)^2}}
\end{aligned}$$

(Ignoring isolated division by 0, which may be repaired by continuity.) Without loss of generality suppose that $z \leq 0$.

$$\ldots \leq Ka\sqrt{\frac{\left(\frac{1-z}{2}\right)^{2K} - \left(\frac{1-z}{2}\right)^{2K-2}\left(\frac{1+z}{2}\right)^2}{\left(\frac{1-z}{2}\right)^2 - \left(\frac{1+z}{2}\right)^2}} = Ka\left(\frac{1-z}{2}\right)^{K-1} \leq Ka.$$

■

Proof of Lemma 2.7. Consider $g \in (\mu \circ D)^{-1}(\vec{0})$. Let $T_n(x)$ denote the Chebyshev polynomials of the first kind and let $U_n(x) = \frac{1}{n+1}\frac{dT_{n+1}(x)}{dx}$ denote the Chebyshev polynomials of the second kind. We will expand $g \in \mathbb{J}_0$ in the basis of the Chebyshev polynomials $\{U_n\}$ to obtain a coefficient bound that is inspired by a classical bound, apparently due to J. E. Littlewood, on the Fourier coefficients of functions in $\mathbb{J}_0$. (We cannot make use of that classical bound but have followed the example of its proof. See Taibleson [38].) Let $c_n = (n+1)\int_{-1}^{1} g(x)U_n(x)\,dx$. We integrate by parts to get

$$c_n = g(1)T_n(1) - g(-1)T_n(-1) - \int_{-1}^{1} T_n(x)\,dD(g)(x) = -\int_{-1}^{1} T_n(x)\,dD(g)(x).$$

Now, for $x \in J$, $|T_n(x)| \le 1$, and $D(g)$ is the difference of two probability measures on J, so $|c_n| \le 2$.

The Chebyshev polynomials $\{U_n\}$ satisfy the orthogonality relations $\frac{2}{\pi}\int_{-1}^{1}\sqrt{1-x^2}\cdot U_m(x)U_n(x)\,dx = \delta_{mn}$, where δ denotes Kronecker delta. Therefore, the inversion formula is

$$g(x) = \sum_{n=0}^{\infty}\frac{2c_n}{\pi(n+1)}\sqrt{1-x^2}\cdot U_n(x) \quad \text{or equivalently} \quad \frac{g(x)}{\sqrt[4]{1-x^2}} = \sum_{n=0}^{\infty}\frac{2c_n}{\pi(n+1)}\sqrt[4]{1-x^2}\cdot U_n(x).$$

The vectors summed on the RHS of the second expression are orthogonal, so

$$\begin{aligned}\int_{-1}^{1}|g(x)|^2\,dx &\le \int_{-1}^{1}\left|\frac{g(x)}{\sqrt[4]{1-x^2}}\right|^2 dx = \sum_{n=0}^{\infty}\left\|\frac{2c_n}{\pi(n+1)}\sqrt[4]{1-x^2}\cdot U_n(x)\right\|_2^2 \\ &= \sum_{n=0}^{\infty}\frac{2c_n^2}{\pi(n+1)^2}\int_{-1}^{1}\frac{2}{\pi}\sqrt{1-x^2}\cdot U_n^2(x)\,dx = \sum_{n=0}^{\infty}\frac{2c_n^2}{\pi(n+1)^2}. \qquad (4)\end{aligned}$$

Now we finally use the assumption that $g \in (\mu \circ D)^{-1}(\vec{0})$. By the same integration by parts formula that we used earlier, we see that c_n is the n-th Chebyshev moment of the signed measure $D(g)$. For $n < K+1$ these moments are linear combinations of the moments $\mathrm{E}\left[x^K\right], \ldots, \mathrm{E}\left[(1-x)^K\right]$. By the assumption that $g \in (\mu \circ D)^{-1}(\vec{0})$, all these moments are equal to 0. In combination with Equation (4), and the power-mean inequality, we obtain: $\left(\int_{-1}^{1}|g(x)|\,dx\right)^2 \le 2\cdot\int_{-1}^{1}|g(x)|^2\,dx \le \sum_{n=K+1}^{\infty}\frac{8}{\pi(n+1)^2} \le \frac{8}{\pi(K+1)}$. ∎

B Proofs from Section 3

Proof of Proposition 3.1.

$$\begin{aligned}M_{i,j} &= \int_{-1}^{1}\left(\frac{1-x}{2}\bar{p}_i + \frac{1+x}{2}\bar{q}_i\right)\cdot\left(\frac{1-x}{2}\bar{p}_j + \frac{1+x}{2}\bar{q}_j\right)d\vartheta \\ &= \frac{1}{4}\left(1-2\mu+\sigma^2+\mu^2\right)\bar{p}_i\bar{p}_j + \frac{1}{4}\left(1+2\mu+\sigma^2+\mu^2\right)\bar{q}_i\bar{q}_j + \frac{1}{4}\left(1-\sigma^2-\mu^2\right)\left(\bar{p}_i\bar{q}_j+\bar{q}_i\bar{p}_j\right) \\ &= \frac{(1-\mu)^2}{4}\bar{p}_i\bar{p}_j + \frac{(1+\mu)^2}{4}\bar{q}_i\bar{q}_j + \frac{1-\mu^2}{4}(\bar{p}_i\bar{q}_j+\bar{q}_i\bar{p}_j) + \frac{\sigma^2}{4}(\bar{p}_i\bar{p}_j+\bar{q}_i\bar{q}_j-\bar{p}_i\bar{q}_j-\bar{q}_i\bar{p}_j) \\ &= r_ir_j + v_iv_j \\ &= \frac{1}{2}\left((r_i-v_i)(r_j-v_j)+(r_i+v_i)(r_j+v_j)\right).\end{aligned}$$

∎

For $s \in [n]$ and $T \subseteq [n]$, let $M_{sT} = \sum_{t\in T} M_{st}$ and let $\tilde{M}_{sT} = \sum_{t\in T}\tilde{M}_{st}$.

Proposition B.1. *Let $s \in [n]$ and $T \subset [n]$. For every $\omega \in \mathbb{N}$ there exists $c > 0$ such that for every $\delta > 0$ and for every $m \geq \frac{c \ln n}{M_{sT}}$, the following event happens with probability at least $1 - n^{-\omega}$.*

$$\left(1 - \sqrt{\frac{c \ln n}{4 \cdot m \cdot M_{sT}}}\right) \cdot M_{sT} \leq \tilde{M}_{sT} \leq \left(1 + \sqrt{\frac{c \ln n}{4 \cdot m \cdot M_{sT}}}\right) \cdot M_{sT}. \tag{5}$$

Proof. Notice that $m \cdot \tilde{M}_{sT}$ is the sum of m iid Bernoulli trials with success probability M_{sT}. Using standard large deviation bounds [3, Corolary A.1.14, page 268],

$$\Pr\left[\left|\tilde{M}_{sT} - M_{sT}\right| > \epsilon M_{sT}\right] < 2e^{-c_\epsilon \cdot m \cdot M_{sT}},$$

where $c_\epsilon = \min\left\{-\ln\left(e^\epsilon (1+\epsilon)^{-(1+\epsilon)}\right), \frac{\epsilon^2}{2}\right\}$. Plugging in $\epsilon = \sqrt{\frac{c \ln n}{4 \cdot m \cdot M_{sT}}}$, and assuming that c is sufficiently large, we get the claimed bounds. ■

Proposition B.2. *For every $\omega \in \mathbb{N}$ there exists $c > 0$ such that for every $\delta > 0$ and for every $m \geq \frac{c}{\delta} \cdot n \ln n$, the following event happens with probability at least $1 - n^{-\omega}$. For all faces s such that $r_s \geq \frac{\delta}{n}$,*

$$\left(1 - \sqrt{(c \cdot n \ln n)/(4 \cdot \delta \cdot m)}\right) \cdot r_s \leq \tilde{r}_s \leq \left(1 + \sqrt{(c \cdot n \ln n)/(4 \cdot \delta \cdot m)}\right) \cdot r_s, \tag{6}$$

and for all faces s such that $r_s < \delta/n$, $\quad 0 \leq \tilde{r}_s < \delta/n + \sqrt{(\delta \cdot c \cdot \ln n)/(4 \cdot n \cdot m)}$.

Proof. Use Lemma B.1 with $T = [n]$ and the union bound over the faces s. ■

Fix $j \in \{1, 2, \ldots, \log(n/\delta)\}$, and let $I = I_j$. Let $\mathcal{E} = \hat{M}_{I\times I} - M_{I\times I}$. For $s \in I$, write $\tilde{r}_s = (1 + \gamma_s) \cdot r_s$. (Assuming Equation (6), $\gamma_s \in \left[-\sqrt{\frac{c \cdot n \ln n}{4 \cdot \delta \cdot m}}, +\sqrt{\frac{c \cdot n \ln n}{4 \cdot \delta \cdot m}}\right]$.) Since $M_{I\times I} = r_I r_I^T + v_I v_I^T$, $\hat{V}_{I\times I} = \hat{M}_{I\times I} - \tilde{r}_I \tilde{r}_I^T = \left(\hat{M}_{I\times I} - M_{I\times I}\right) + \left(M_{I\times I} - r_I r_I^T\right) + \left(r_I r_I^T - \tilde{r}_I \tilde{r}_I^T\right) = \mathcal{E} + v_I v_I^T - \mathcal{T}$, where $\mathcal{T}_{st} = \left(\gamma_s + \gamma_t + \gamma_s \gamma_t\right) r_s r_t$. The following lemma is a consequence of the eigenvalue concentration bounds of [2, 41].

Lemma B.3. *For every $\omega \in \mathbb{N}$ there exist $\alpha, c > 0$ such that for every $S \geq \ln^3 n$, if $\mathrm{E}[\hat{m}] = c \cdot |I| \cdot S^2$, then with probability at least $1 - n^{-\omega}$,*

$$|\lambda_1(\mathcal{E})| \leq \frac{\alpha}{|I| \cdot S}. \tag{7}$$

Proof. We may assume that with the desired probability, for every $s, t \in I$, $M_{st} = r_s r_t + v_s v_t \leq 2 r_s r_t < 5 \tilde{r}_s \tilde{r}_t \leq \frac{20}{|I|^2}$, as this follows from Equation (6), taking a sufficiently large c.

Consider the random symmetric matrix $A = \hat{m} \cdot \left(\hat{M}_{I\times I} - M_{I\times I}\right)$. The entries A_{st}, $1 \leq s \leq t \leq |I|$, are independent random variables as $\hat{m}$ is a Poisson random variable. Also, for every $s, t \in [I]$, $\mathrm{E}[A_{st}] = 0$, and for every $\omega \in \mathbb{N}$ there exists $\kappa > 0$ such that $\Pr[|A_{st}| > \kappa \ln n] < n^{-\omega}$. Thus, with the desired probability, for some $\kappa > 0$, for every $s, t \in [I]$, $|A_{st}| \leq \kappa \ln n$. Also, with the desired probability, $\hat{m} \geq \frac{1}{2} \mathrm{E}[\hat{m}]$.

Define a matrix B as follows. Let $\overline{A}$ be the truncated version of A with $\overline{A}_{st} = 0$ if $\hat{m} < \frac{1}{2} \mathrm{E}[\hat{m}]$ and $\min\{A_{st}, \kappa \ln n\}$ otherwise. Put $B = \overline{A} - \mathrm{E}[\overline{A}]$. Notice that $\lambda_1(\mathrm{E}[\overline{A}]) \to 0$ as $c, \kappa \to \infty$.

Now, for every $s, t \in I$, $\mathrm{E}[B_{st}] = 0$. Furthermore, Notice that $\hat{m} \cdot \hat{M}_{st}$ is distributed Poisson with expectation (and variance) $\mathrm{E}[\hat{m}] \cdot M_{st}$. Therefore,

$$\mathrm{Var}[B_{st}] \leq \mathrm{Var}[A_{st}] = \mathrm{Var}[\hat{m} \cdot (\hat{M}_{st} - M_{st})] = \mathrm{E}[\hat{m}] \cdot M_{st} \leq \left(c \cdot |I| \cdot S^2\right) \cdot \left(\frac{20}{|I|^2}\right) = \frac{20c \cdot S^2}{|I|}.$$

We now use the following bound due to Alon et al. [2] and Vu [41]. If B is a random matrix as above then there exists a constant $\beta > 0$ such that with the desired probability

$$\lambda_1(B) \leq \beta \cdot \left(\sqrt{\frac{20c \cdot S^2}{|I|}} \cdot \sqrt{|I|} + \sqrt{2\kappa \ln n \cdot \sqrt{\frac{20c \cdot S^2}{|I|}} \cdot \sqrt[4]{|I|} \cdot \ln |I|} \right) = O(S).$$

With the desired probability, $\hat{m} \geq \frac{1}{2} \operatorname{E}[\hat{m}]$, $A = \overline{A}$, and $\lambda_1(B) = O(S)$. Hence, for some constant $\alpha > 0$, we have $|\lambda_1(\mathcal{E})| = \frac{1}{\hat{m}} \cdot \lambda_1(A) \leq \frac{2}{\operatorname{E}[\hat{m}]} \cdot \left(\lambda_1(B) + \lambda_1(\operatorname{E}[\overline{A}])\right) \leq \frac{\alpha}{|I| \cdot S}$. ■

Proposition B.4. $|\lambda_1(\mathcal{T})| \leq 3 \cdot \max_{s \in I} |\gamma_s| \cdot \sum_{s \in I} r_s^2$.

Proof. Let $\rho \in \mathbb{R}^I$ be given by $\rho_s = \gamma_s r_s$. Then, $\mathcal{T} = \rho r_I^T + r_I \rho^T + \rho \rho^T$. Therefore,

$$\begin{aligned} |\lambda_1(\mathcal{T})| &\leq \left|\lambda_1(\rho r_I^T)\right| + \left|\lambda_1(r_I \rho^T)\right| + \left|\lambda_1(\rho \rho^T)\right| = \left|\sum_{s \in I} \gamma_s r_s^2\right| + \left|\sum_{s \in I} \gamma_s^2 r_s^2\right| \\ &\leq 3 \left|\sum_{s \in I} \gamma_s r_s^2\right| \leq 3 \cdot \max_{s \in I} |\gamma_s| \cdot \sum_{s \in I} r_s^2, \end{aligned}$$

as stipulated. ■

Lemma B.5. *Assuming Equations* (6) *and* (7)*, there is a constant $\beta > 0$ such that*

$$\|v_I\|_2^2 - \left(1/S + \sqrt{(n \ln n)/(\delta \cdot m)}\right) \cdot \beta/|I| \leq \lambda(I) \leq \|v_I\|_2^2 + \left(1/S + \sqrt{(n \ln n)/(\delta \cdot m)}\right) \cdot \beta/|I|; \quad (8)$$

$$\left(\hat{v}_I^T v_I\right)^2 \geq \|v_I\|_2^2 - \left(1/S + \sqrt{(n \ln n)/(\delta \cdot m)}\right) \cdot \beta/|I|; \quad \textit{and} \quad (9)$$

if $\hat{v}_I^T v_I \geq 0$ *then* $\left\| \sqrt{\lambda(I)} \cdot \hat{v}_I - v_I \right\|_1^2 \leq |I| \cdot \left\| \sqrt{\lambda(I)} \cdot \hat{v}_I - v_I \right\|_2^2 \leq \beta \cdot \left(1/S + \sqrt{(n \ln n)/(\delta \cdot m)}\right)$. (10)

Proof. Notice that

$$\lambda(I) = \hat{v}_I^T \hat{V}_{I \times I} \hat{v}_I \leq \hat{v}_I^T V_{I \times I} \hat{v}_I + \max_{y \neq 0} \frac{y^T \mathcal{E} y}{y^T y} + \max_{y \neq 0} \frac{y^T \mathcal{T} y}{y^T y} = \left(\hat{v}_I^T v_I\right)^2 + |\lambda_1(\mathcal{E})| + |\lambda_1(\mathcal{T})|.$$

On the other hand, $\lambda(I) = \max_{y \neq 0} \frac{y^T \hat{V}_{I \times I} y}{y^T y} \geq v_I^T v_I - |\lambda_1(\mathcal{E})| - |\lambda_1(\mathcal{T})|$. Combining the two inequalities, we get that $\left(\hat{v}_I^T v_I\right)^2 \geq \|v_I\|_2^2 - 2|\lambda_1(\mathcal{E})| - 2|\lambda_1(\mathcal{T})|$. By Equation (7), for some constant $\beta_1 > 0$, we have $2|\lambda_1(\mathcal{E})| \leq \frac{\beta_1}{|I|} \cdot \frac{1}{S}$. Moreover, by Equation (6), $\max_{s \in I} |\gamma_s| = O\left(\sqrt{\frac{n \ln n}{\delta \cdot m}}\right)$. Furthermore

$$\sum_{s \in I} r_s^2 \leq 4 \cdot \sum_{s \in I} \tilde{r}_s^2 \leq \frac{8}{|I|} \cdot \sum_{s \in I} \tilde{r}_s \leq \frac{8}{|I|}.$$

Therefore, by Proposition B.4 there is a constant $\beta_2 > 0$ such that $2|\lambda_1(\mathcal{T})| \leq \frac{\beta_2}{|I|} \cdot \sqrt{\frac{n \ln n}{4 \cdot \delta \cdot m}}$. Thus we get Equation (8) (as $\left(\hat{v}_I^T v_I\right)^2 \leq \|v_I\|_2^2$) and Equation (9). Using the fact that $\sqrt{1-\epsilon} \geq 1 - 2\epsilon$ for all $\epsilon \in [0,1]$, we get that for a constant $\beta_3 > 0$,

$$\left|\hat{v}_I^T v_I\right| \geq \|v_I\|_2 - \frac{\beta_3}{\|v_I\|_2 \cdot |I|} \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right).$$

Let $\hat{v}'_I = \|v_I\|_2 \cdot \hat{v}_I$. Assuming that $\hat{v}_I^T v_I \geq 0$, we get that

$$\left\|\hat{v}'_I - v_I\right\|_2^2 = 2 \cdot \|v_I\|_2^2 - 2 \cdot \|v_I\|_2 \cdot \hat{v}_I^T v_I \leq \frac{2\beta_3}{|I|} \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right).$$

Also, by Equation (8), for a constant $\beta_4 > 0$,

$$\left\|\sqrt{\lambda(I)} \cdot \hat{v}_I - \hat{v}'_I\right\|_2^2 = \left(\sqrt{\lambda(I)} - \|v_I\|_2\right)^2 \leq \left|\lambda(I) - \|v_I\|_2^2\right| \leq \frac{\beta_4}{|I|} \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right).$$

Thus, for a constant $\beta_5 > 0$, we have

$$\begin{aligned}
\left\|\sqrt{\lambda(I)} \cdot \hat{v}_I - v_I\right\|_1^2 &\leq |I| \cdot \left\|\sqrt{\lambda(I)} \cdot \hat{v}_I - v_I\right\|_2^2 \\
&\leq |I| \cdot \left(\left\|\hat{v}'_I - v_I\right\|_2^2 + \left\|\sqrt{\lambda(I)} \cdot \hat{v}_I - \hat{v}'_I\right\|_2^2 + 2 \cdot \left\|\hat{v}'_I - v_I\right\|_2 \cdot \|\lambda(I) \cdot \hat{v}_I - v_I\|_2\right) \\
&\leq \beta_5 \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right),
\end{aligned}$$

showing Equation (10). ■

Given our bounds on δ, ϵ, S, m, and $c > 0$ sufficiently large, Equations (6), (7), (8), (9), and (10) hold with high probability.

Lemma B.6. *$J \neq \emptyset$ and for every $j \in J$, $\tilde{v}_{I_j}^T v_{I_j} > 0$.*

Proof. We first show the first part of the claim. By Equation (6), for every s such that $\tilde{r}_s \geq \frac{\delta}{n}$, $r_s \leq 2\tilde{r}_s$, and for every s such that $\tilde{r}_s < \frac{\delta}{n}$, $r_s < \frac{2\delta}{n}$ Put $J_0 = \{j \in \mathbb{N} : \ j > \log(n/\delta)\}$; $J_{<\epsilon} = \{j \in \mathbb{N} : \ j \leq \log(n/\delta) \ \wedge \ \sum_{s \in I_j} \tilde{r}_s < \epsilon\}$; and $J_{\geq\epsilon} = \{j \in \mathbb{N} : \ j \leq \log(n/\delta) \ \wedge \ \sum_{s \in I_j} \tilde{r}_s \geq \epsilon\}$. We have that

$$\sum_{j \in J_0} \sum_{s \in I_j} |v_s| \leq \sum_{j \in J_0} \sum_{s \in I_j} r_s < 2\delta.$$

Also,

$$\sum_{j \in J_{<\epsilon}} \sum_{s \in I_j} |v_s| \leq \sum_{j \in J_{<\epsilon}} \sum_{s \in I_j} r_s \leq 2 \cdot \sum_{j \in J_{<\epsilon}} \sum_{s \in I_j} \tilde{r}_s < 2\delta.$$

Assume for contradiction that for all $j \in J_{\geq\epsilon}$, $\sum_{s \in I_j} |v_s| < \epsilon$. Then,

$$\sum_{s=1}^{n} |v_s| = \sum_{j \in J_0} \sum_{s \in I_j} |v_s| + \sum_{j \in J_{<\epsilon}} \sum_{s \in I_j} |v_s| + \sum_{j \in J_{\geq\epsilon}} \sum_{s \in I_j} |v_s| < 2\delta + 2\delta + \delta \leq \zeta,$$

in contradiction to the definition of ζ. By Equation (8), there is a constant β such that for every $j \in J_{\geq\epsilon}$,

$$\lambda(I_j) \geq \frac{\|v_{I_j}\|_1^2}{|I_j|} - \frac{\beta}{|I_j|} \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right) \geq \frac{\|v_{I_j}\|_1^2}{2|I_j|}.$$

As we have shown that there exists $j \in J_{\geq\epsilon}$ such that $\|v_{I_j}\|_1 \geq \epsilon$, this completes the proof.

We now proceed with the second part of the claim. Consider $j \in J$ such that $\hat{v}_{I_j}^T v_{I_j} > 0$ and $\|\hat{v}_{\hat{T}_j^+}\|_2^2 \geq \frac{1}{2}$. (The other cases have parallel proofs.) Let $B_j = \hat{T}_j^+ \triangle T_j^+$ and let $A = I_j \setminus B_j$. Notice that by Equation (9),

$$\|v_A\|_2^2 \geq \left(\hat{v}_A^T v_A\right)^2 \geq \left(\hat{v}_{I_j}^T v_{I_j}\right)^2 \geq \|v_{I_j}\|_2^2 - \frac{\beta}{|I_j|} \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right),$$

and therefore

$$\|v_{B_j}\|_1^2 \leq |I_j| \cdot \|v_{B_j}\|_2^2 \leq \beta \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right).$$

Also, as for every $s \in I_j$, $|v_s| \leq r_s \leq 2\tilde{r}_s \leq \frac{4}{|I_j|}$, using Lemma B.5 and previous arguments,

$$\begin{aligned}
\|v_{\hat{T}_j^+}\|_1 &\geq \frac{|I_j|}{4} \cdot \|v_{\hat{T}_j^+}\|_2^2 \\
&\geq \frac{|I_j|}{4} \cdot \left(\|\sqrt{\lambda(I_j)} \cdot \hat{v}_{\hat{T}_j^+}\|_2^2 - 2 \cdot \|\sqrt{\lambda(I_j)} \cdot \hat{v}_{\hat{T}_j^+}\|_2 \cdot \|\sqrt{\lambda(I_j)} \cdot \hat{v}_{\hat{T}_j^+} - v_{\hat{T}_j^+}\|_2\right) \\
&\geq \frac{|I_j|}{4} \cdot \left(\|\sqrt{\lambda(I_j)} \cdot \hat{v}_{\hat{T}_j^+}\|_2^2 - 2 \cdot \|\sqrt{\lambda(I_j)} \cdot \hat{v}_{I_j}\|_2 \cdot \|\sqrt{\lambda(I_j)} \cdot \hat{v}_{I_j} - v_{I_j}\|_2\right) \\
&\geq \frac{|I_j|}{4} \cdot \left(\frac{\lambda(I_j)}{2} - 2 \cdot \sqrt{\lambda(I_j)} \cdot \sqrt{\frac{\beta}{|I_j|} \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right)}\right) \\
&\geq \frac{\epsilon^2}{32}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
e_j &= \sum_{s \in \hat{T}_j^+} \sum_{t \in \hat{T}_{j_0}^+} (\tilde{M}_{st} - \tilde{r}_s \tilde{r}_t) \geq \sum_{s \in \hat{T}_j^+} \sum_{t \in \hat{T}_{j_0}^+} (M_{st} - r_s r_t) - O\left(\sqrt{\frac{n \ln n}{\delta \cdot m}}\right) = \sum_{s \in \hat{T}_j^+} \sum_{t \in \hat{T}_{j_0}^+} v_s v_t - O\left(\sqrt{\frac{n \ln n}{\delta \cdot m}}\right) \\
&\geq \sum_{s \in \hat{T}_j^+} \sum_{t \in \hat{T}_{j_0}^+} |v_s| \cdot |v_t| - 2 \cdot \sum_{s \in B_j} \sum_{t \in \hat{T}_{j_0}^+} |v_s| \cdot |v_t| - 2 \cdot \sum_{s \in \hat{T}_j^+} \sum_{t \in B_{j_0}} |v_s| \cdot |v_t| - O\left(\sqrt{\frac{n \ln n}{\delta \cdot m}}\right) \\
&= \|v_{\hat{T}_j^+}\|_1 \cdot \left(\frac{1}{2} \cdot \|v_{\hat{T}_{j_0}}\|_1 - 2 \cdot \|v_{B_{j_0}}\|_1\right) + \|v_{\hat{T}_{j_0}^+}\|_1 \cdot \left(\frac{1}{2} \cdot \|v_{T_j^+}\|_1 - \|v_{B_j}\|_1\right) - O\left(\sqrt{\frac{n \ln n}{\delta \cdot m}}\right) \\
&\geq \left(\|v_{\hat{T}_j^+}\|_1 + \|v_{\hat{T}_{j_0}^+}\|_1\right) \cdot \left(\frac{\epsilon^2}{32} - \sqrt{\beta \cdot \left(\frac{1}{S} + \sqrt{\frac{n \ln n}{\delta \cdot m}}\right)}\right) - O\left(\sqrt{\frac{n \ln n}{\delta \cdot m}}\right) > 0.
\end{aligned}$$

This completes the proof. ■

Proof of Theorem 3.2. Put $\xi = \log(n/\delta) \cdot \sqrt{\beta \cdot \left(1/S + \sqrt{(n \ln n)/(\delta \cdot m)}\right)}$. Let $B_j = \hat{T}_j^+ \triangle T_j^+$. Following the proof of Lemma B.6 (in Appendix B), $\sum_{s \in \tilde{T}^+} v_s \geq \sum_{s \in T^+} v_s - \sum_{j \notin J} \sum_{s \in T_j^+} v_s - \sum_{j \in J} \sum_{s \in B_j} |v_s| \geq \zeta - 5\delta - \xi > \frac{\zeta}{2}$. Similarly, $-\sum_{s \in \tilde{T}^-} v_s > \frac{\zeta}{2}$. By Equation (10), $\left|\sum_{s \in \tilde{T}^+} (v_s - \tilde{v}_s)\right| \leq \sum_{s \in \tilde{T}^+} |v_s - \tilde{v}_s| \leq \sum_{j \in J} \|v_{I_j} - \tilde{v}_{I_j}\|_1 \leq \xi$, and $\left|\sum_{s \in \tilde{T}^-} (v_s - \tilde{v}_s)\right| \leq \xi$. Thus, $\sum_{s \in \tilde{T}^+} \tilde{v}_s \geq \zeta - 5\delta - 2\xi > \frac{\zeta}{2}$, and $-\sum_{s \in \tilde{T}^-} \tilde{v}_s > \frac{\zeta}{2}$. Set $p = \frac{\sum_{s \in \tilde{T}^+} (r_s + \alpha \cdot v_s)}{\sum_{s \in \tilde{T}^+ \cup \tilde{T}^-} (r_s + \alpha \cdot v_s)}$ and $\varphi(\alpha) = \frac{p \cdot \sum_{s \in \tilde{T}^-} \tilde{r}_s - (1-p) \cdot \sum_{s \in \tilde{T}^+} \tilde{r}_s}{(1-p) \cdot \sum_{s \in \tilde{T}^+} \tilde{v}_s - p \cdot \sum_{s \in \tilde{T}^-} \tilde{v}_s}$. Notice that $0 \leq p \leq 1$ and $\varphi(\alpha)$ is well-defined as we've shown that the denominator in its expression is non-zero. This verifies Equation (2).

Also notice that $\alpha = \frac{p\cdot\sum_{s\in\tilde{T}^-} r_s-(1-p)\cdot\sum_{s\in\tilde{T}^+} r_s}{(1-p)\cdot\sum_{s\in\tilde{T}^+} v_s-p\cdot\sum_{s\in\tilde{T}^-} v_s}$. (Here, too, we've shown that the denominator is non-zero.) We now upper-bound $|\alpha-\varphi(\alpha)|$ using the equality $\frac{A}{B}-\frac{C}{D} = \frac{A-C}{D}+\frac{(D-B)\cdot A}{B\cdot D}$. By Equation (6), $|A-C| \le p\cdot\sum_{s\in\tilde{T}^-}|r_s-\tilde{r}_s|+(1-p)\cdot\sum_{s\in\tilde{T}^+}|r_s-\tilde{r}_s| \le \sqrt{\frac{c\cdot n\ln n}{4\cdot\delta\cdot m}}$. As argued above, $B=(1-p)\cdot\sum_{s\in\tilde{T}^+} v_s - p\cdot\sum_{s\in\tilde{T}^-} v_s \ge \frac{\zeta}{2}$, and $D=(1-p)\cdot\sum_{s\in\tilde{T}^+}\tilde{v}_s - p\cdot\sum_{s\in\tilde{T}^-}\tilde{v}_s \ge \frac{\zeta}{2}$. Also as argued above, $|D-B| \le (1-p)\cdot\left|\sum_{s\in\tilde{T}^+}(\tilde{v}_s-v_s)\right|+p\cdot\left|\sum_{s\in\tilde{T}^-}(\tilde{v}_s-v_s)\right| \le \xi$. Finally, $|A| \le p\cdot\sum_{s\in\tilde{T}^-} r_s+(1-p)\cdot\sum_{s\in\tilde{T}^+} r_s \le 1$. Therefore, Equation (1) follows from $|\alpha-\varphi(\alpha)| \le \frac{2}{\zeta}\cdot\sqrt{\frac{c\cdot n\ln n}{4\cdot\delta\cdot m}}+\frac{4}{\zeta^2}\cdot\log(n/\delta)\cdot\sqrt{\beta\cdot\left(1/S+\sqrt{(n\ln n)/(\delta\cdot m)}\right)}$.

Let $I_0=\{s\in[n]:\ r_s<\frac{\delta}{n}\}$. Using Equation (6), $\|r-\tilde{r}\|_1 = \sum_{s\in I_0}|r_s-\tilde{r}_s|+\sum_{s\notin I_0}|r_s-\tilde{r}_s| \le 2\delta+\sqrt{\frac{c\cdot n\ln n}{4\cdot\delta\cdot m}}$. As the L_1 distance between any two points in the simplex is at most 2 and $\|v\|_1 = 2\zeta$, it must be that $|\alpha| \le \frac{1}{\zeta}$. Similar to the argument above, $\|v-\tilde{v}\|_1 = \sum_{j\in J}\|v_{I_j}-\tilde{v}_{I_j}\|_1+\sum_{j\notin J}\|v_{I_j}\|_1 \le \xi+5\delta$. Also, $\|\tilde{v}\|_1 = \sum_{j\in J}\|\tilde{v}_{I_j}\|_1 \le \sum_{j\in J}\|v_{I_j}\|_1+\sum_{j\in J}\|\tilde{v}_{I_j}-v_{I_j}\|_1 \le 1+\xi$. Using Equation (1), as $\|(r+\alpha\cdot v)-(\tilde{r}+\varphi(\alpha)\cdot\tilde{v})\|_1 \le \|r-\tilde{r}\|_1+|\alpha|\cdot\|v-\tilde{v}\|_1+|\alpha-\varphi(\alpha)|\cdot\|\tilde{v}\|_1$, we get Equation (3). ■

C Transportation distance for discrete distributions

Definition C.1. Let (ϑ_1,α_1) and (ϑ_2,α_2) represent k-spike and ℓ-spike distributions respectively. The transportation distance between these two distributions, denoted by $\mathrm{Tran}(\vartheta_1,\alpha_1;\vartheta_2,\alpha_2)$, is the optimum value of the following minimum-cost flow linear program:

$$\min \sum_{i=1}^{k}\sum_{j=1}^{\ell} x_{ij}|\alpha_{1i}-\alpha_{2j}| \qquad \text{subject to}$$

$$\sum_{j=1}^{\ell} x_{ij} = \vartheta_{1i} \quad \forall i=1,\ldots,k; \qquad \sum_{i=1}^{k} x_{ij} = \vartheta_{2j} \quad \forall j=1,\ldots,\ell; \qquad x_{ij}\ge 0 \quad \forall i,j.$$

This is the *transportation LP*, one of the early examples of a linear program [12]. Note that $\sum_{ij} x_{ij} = 1$.