JMLR: Workshop and Conference Proceedings vol 35:1–30, 2014

The More, the Merrier: the Blessing of Dimensionality for Learning Large Gaussian Mixtures

Joseph Anderson ANDEJOSE@CSE.OHIO-STATE.EDU
Mikhail Belkin MBELKIN@CSE.OHIO-STATE.EDU
Department of Computer Science and Engineering, the Ohio State University

Navin Goyal NAVINGO@MICROSOFT.COM
Microsoft Research India

Luis Rademacher LRADEMAC@CSE.OHIO-STATE.EDU
James Voss VOSSJ@CSE.OHIO-STATE.EDU
Department of Computer Science and Engineering, the Ohio State University

Abstract

In this paper we show that very large mixtures of Gaussians are efficiently learnable in high dimension. More precisely, we prove that a mixture with known identical covariance matrices whose number of components is a polynomial of any fixed degree in the dimension n is polynomially learnable as long as a certain non-degeneracy condition on the means is satisfied. It turns out that this condition is generic in the sense of smoothed complexity, as soon as the dimensionality of the space is high enough. Moreover, we prove that no such condition can possibly exist in low dimension and the problem of learning the parameters is generically hard. In contrast, much of the existing work on Gaussian Mixtures relies on low-dimensional projections and thus hits an artificial barrier.

Our main result on mixture recovery relies on a new "Poissonization"-based technique, which transforms a mixture of Gaussians to a linear map of a product distribution. The problem of learning this map can be efficiently solved using some recent results on tensor decompositions and Independent Component Analysis (ICA), thus giving an algorithm for recovering the mixture. In addition, we combine our low-dimensional hardness results for Gaussian mixtures with Poissonization to show how to embed difficult instances of low-dimensional Gaussian mixtures into the ICA setting, thus establishing exponential information-theoretic lower bounds for underdetermined ICA in low dimension. To the best of our knowledge, this is the first such result in the literature.

In addition to contributing to the problem of Gaussian mixture learning, we believe that this work is among the first steps toward better understanding the rare phenomenon of the "blessing of dimensionality" in the computational aspects of statistical inference.

Keywords: Gaussian mixture models, tensor methods, blessing of dimensionality, smoothed analysis, Independent Component Analysis

1. Introduction

The question of recovering a probability distribution from a finite set of samples is one of the most fundamental questions of statistical inference. While classically such problems have been considered in low dimension, more recently inference in high dimension has drawn significant attention in statistics and computer science literature.

In particular, an active line of investigation in theoretical computer science has dealt with the question of learning a Gaussian Mixture Model in high dimension. This line of work was

started in Dasgupta (1999) where the first algorithm to recover parameters using a number of samples polynomial in the dimension was presented. The method relied on random projections to a low dimensional space and required certain separation conditions for the means of the Gaussians. Significant work was done in order to weaken the separation conditions and to generalize the result (see e.g., Dasgupta and Schulman (2000); Arora and Kannan (2001); Vempala and Wang (2002); Achlioptas and McSherry (2005); Feldman et al. (2006)). Much of this work has polynomial sample and time complexity but requires strong separation conditions on the Gaussian components. A completion of the attempts to weaken the separation conditions was achieved in Belkin and Sinha (2010) and Moitra and Valiant (2010), where it was shown that arbitrarily small separation was sufficient for learning a general mixture with a fixed number of components in polynomial time. Moreover, a one-dimensional example given in Moitra and Valiant (2010) showed that an exponential dependence on the number of components was unavoidable unless strong separation requirements were imposed. Thus the question of polynomial learnability appeared to be settled. It is worth noting that while quite different in many aspects, all of these papers used a general scheme similar to that in the original work Dasgupta and Schulman (2000) by reducing high-dimensional inference to a small number of low-dimensional problems through appropriate projections.

However, a surprising result was recently proved in Hsu and Kakade (2013). The authors showed that a mixture of d Gaussians in dimension d could be learned using a polynomial number of samples, assuming a non-degeneracy condition on the configuration of the means. The result in Hsu and Kakade (2013) is inherently high-dimensional as that condition is never satisfied when the means belong to a lower-dimensional space. Thus the problem of learning a mixture gets progressively computationally easier as the dimension increases, a "blessing of dimensionality!" It is important to note that this was quite different from much of the previous work, which had primarily used projections to lower-dimension spaces.

Still, there remained a large gap between the worst case impossibility of efficiently learning more than a fixed number of Gaussians in low dimension and the situation when the number of components is equal to the dimension. Moreover, it was not completely clear whether the underlying problem was genuinely easier in high dimension or our algorithms in low dimension were suboptimal. The one-dimensional example in Moitra and Valiant (2010) cannot answer this question as it is a specific worst-case scenario, which can be potentially ruled out by some genericity condition.

In our paper we take a step to eliminate this gap by showing that even very large mixtures of Gaussians can be polynomially learned. More precisely, we show that a mixture of m Gaussians with equal known covariance can be polynomially learned as long as m is bounded from above by a polynomial of the dimension n and a certain more complex non-degeneracy condition for the means is satisfied. We show that if n is high enough, these non-degeneracy conditions are generic in the smoothed complexity sense. Thus for any fixed d, $O(n^d)$ generic Gaussians can be polynomially learned in dimension n.

Further, we prove that no such condition can exist in low dimension. A measure of non-degeneracy must be monotone in the sense that adding Gaussian components must make the condition number worse. However, we show that for k^2 points uniformly sampled from $[0, 1]$ there are (with high probability) two mixtures of unit Gaussians with means on non-intersecting subsets of these points, whose L^1 distance is $O^*(e^{-k})$ and which are thus not polynomially identifiable. More generally, in dimension n the distance becomes $O^*(\exp(-\sqrt[n]{k}))$. That is, the conditioning improves as the dimension increases, which is consistent with our algorithmic results.

To summarize, our contributions are as follows:

(1) We show that for any q, a mixture of n^q Gaussians in dimension n can be learned in time and number of samples polynomial in n and a certain "condition number" σ. For sufficiently high dimension, this results in an algorithm polynomial from the smoothed analysis point of view (Theorem 1). To do that we provide smoothed analysis of the condition number using certain results from Rudelson and Vershynin (2009) and anti-concentration inequalities. The main technical ingredient of the algorithm is a new "Poissonization" technique to reduce Gaussian mixture estimation to a problem of recovering a linear map of a product distribution known as underdetermined Independent Component Analysis (ICA). We combine this with the recent work on efficient algorithms for underdetermined ICA from Goyal et al. (2013) to obtain the necessary bounds.
(2) We show that in low dimension polynomial identifiability fails in a certain generic sense (see Theorem 3). Thus the efficiency of our main algorithm is truly a consequence of the "blessing of dimensionality" and no comparable algorithm exists in low dimension. The analysis is based on results from approximation theory and Reproducing Kernel Hilbert Spaces.

Moreover, we combine the approximation theory results with the Poissonization-based technique to show how to embed difficult instances of low-dimensional Gaussian mixtures into the ICA setting, thus establishing exponential information-theoretic lower bounds for underdetermined ICA in low dimension. To the best of our knowledge, this is the first such result in the literature.

We discuss our main contributions more formally now. The notion of Khatri–Rao power $A^{\odot d}$ of a matrix A is defined in Section 2.

Theorem 1 (Learning a GMM with Known Identical Covariance) *Suppose $m \geq n$ and let $\epsilon, \delta > 0$. Let $w_1\mathcal{N}(\mu_1, \Sigma) + \cdots + w_m\mathcal{N}(\mu_m, \Sigma)$ be an n-dimensional GMM, i.e. $\mu_i \in \mathbb{R}^n$, $w_i > 0$, and $\Sigma \in \mathbb{R}^{n\times n}$. Let B be the $n \times m$ matrix whose i^{th} column is $\mu_i / \|\mu_i\|$. If there exists $d \in \mathbb{N}$ so that $\sigma_m\left(B^{\odot d}\right) > 0$, then Algorithm 2 recovers each μ_i to within ϵ accuracy with probability $1 - \delta$. Its sample and time complexity are at most*

$$\mathrm{poly}\left(m^{d^2}, \sigma^{d^2}, u^{d^2}, w^{d^2}, d^{d^2}, r^{d^2}, 1/\epsilon, 1/\delta, 1/b, \log^{d^2}\left(1/(b\epsilon\delta)\right)\right)$$

where $w \geq \max_i(w_i)/\min_i(w_i)$, $u \geq \max_i \|\mu_i\|$, $r \geq \left(\max_i \|\mu_i\| + 1\right)/\left(\min_i \|\mu_i\|\right)$, $0 < b \leq \sigma_m(B^{\odot d})$ are bounds provided to the algorithm, and $\sigma = \sqrt{\lambda_{\max}(\Sigma)}$.

We note that the requirement that $m \geq n$ is due to the invocation of Theorem 1.3 of Goyal et al. (2013); it should not be difficult, however, to adapt the algorithm to use a method similar to that of Hsu and Kakade (2013) to handle the case where $m < n$.

Given that the means have been estimated, the weights can be recovered using the tensor structure of higher order cumulants (see Section 2 for the definition of cumulants). This is shown in Appendix I.

We show that $\sigma_{\min}(A^{\odot d})$ is large in the smoothed analysis sense, namely, if we start with a base matrix A and perturb each entry randomly to get $\tilde{A}$, then $\sigma_{\min}(\tilde{A}^{\odot d})$ is likely to be large:

Theorem 2 *For $n > 1$, let $M \in \mathbb{R}^{n\times\binom{n}{2}}$ be an arbitrary matrix. Let $N \in \mathbb{R}^{n\times\binom{n}{2}}$ be a randomly sampled matrix with each entry iid from $\mathcal{N}(0, \sigma^2)$, for $\sigma > 0$. Then, for some absolute constant C, $\Pr\left(\sigma_{\min}((M+N)^{\odot 2}) \leq \sigma^2/n^7\right) \leq 2C/n$.*

We point out the simultaneous and independent work of Bhaskara et al. (2014), where the authors prove learnability results related to our Theorems 1 and 2. We now provide a comparison. The results in Bhaskara et al. (2014), which are based on tensor decompositions, are stronger in that they can learn mixtures of axis-aligned Gaussians (with non-identical covariance matrices) without

requiring to know the covariance matrices in advance. Their results hold under a smoothed analysis setting similar to ours. To learn a mixture of roughly $n^{\ell/2}$ Gaussians up to an accuracy of ϵ their algorithm has running time and sample complexity $\mathrm{poly}_\ell(n, 1/\epsilon, 1/\rho)$ and succeeds with probability at least $1 - \exp(-Cn^{1/3^\ell})$, where the means are perturbed by adding an n-dimensional Gaussian from $\mathcal{N}(0, I_n\rho^2/n)$. On the one hand, the success probability of their algorithm is much better (as a function of n, exponentially close to 1 as opposed to polynomially close to 1, as in our result). On the other hand, this comes at a high price in terms of the running time and sample complexity: The polynomial $\mathrm{poly}_\ell(n, 1/\epsilon, 1/\rho)$ above has degree exponential in ℓ, unlike the degree of our bound which is polynomial in ℓ. Thus, in this respect, the two results can be regarded as incomparable points on an error vs running time (and sample complexity) trade-off curve. Our result is based on a reduction from learning GMMs to ICA which could be of independent interest given that both problems are extensively studied in somewhat disjoint communities. Moreover, our analysis can be used in the reverse direction to obtain hardness results for ICA.

The technique of Poissonization is, to the best of our knowledge, new in the GMM setting, though we note that it has been previously applied in computational learning. For instance, in Valiant and Valiant (2013), it has been used for the estimation of properties of discrete probability distributions such as the support size and the entropy.

Finally, in Section 6 we show that in low dimension the situation is very different from the high-dimensional generic efficiency given by Theorems 1 and 2: The problem is generically hard. More precisely, we show:

Theorem 3 *Let X be a set of $4k^2$ points uniformly sampled from $[0,1]^n$. Then with high probability there exist two mixtures with equal number of unit Gaussians p, q centered on disjoint subsets of X, such that, for some $C > 0$, $\|p - q\|_{L^1(\mathbb{R}^n)} < \exp\big(-C\,(k/\log k)^{1/n}\big)$.*

Here we would like to note that the assumption that X is random is convenient as it provides a natural model for *genericity*, guarantees (with high probability) small *fill* (Section 6) and also ensures that the means of the Gaussian mixture components of p and q are not too close. In particular, it is not difficult to verify that with high probability any pair of the random means are at least $1/\mathrm{poly}(k)$ separated. However the randomness assumption is not essential. In fact, the above theorem will hold for an arbitrary set of points with sufficiently small fill (see Section 6).

Combining the above lower bound with our reduction provides a similar lower bound for ICA; see a discussion on the connection with ICA below. Our lower bound gives an information-theoretic barrier. This is in contrast to conjectured computational barriers that arise in related settings based on the noisy parity problem (see Hsu and Kakade (2013) for pointers). The only previous information-theoretic lower bound for learning GMMs we are aware of is due to Moitra and Valiant (2010) and holds for two specially designed one-dimensional mixtures.

Connection with ICA. A key observation of Hsu and Kakade (2013) is that methods based on the higher order statistics used in Independent Component Analysis (ICA) can be adapted to the setting of learning a Gaussian Mixture Model. In ICA, samples are of the form $X = \sum_{i=1}^m A_i S_i$ where the latent random variables S_i are independent, and the fixed and unknown column vectors A_i give the directions in which each signal S_i acts. The goal is to recover the vectors A_i up to inherent ambiguities. The ICA problem is typically posed when m is at most the dimensionality of the observed space (the "fully determined" setting), as recovery of the directions A_i then allows one to demix the latent signals. The case where the number of latent source signals exceeds the dimensionality of the observed signal X is the *underdetermined ICA* setting.[1] Two well-known

1. See (Comon and Jutten, 2010, Chapter 9) for a recent account of algorithms for underdetermined ICA.

algorithms for underdetermined ICA are given in Cardoso (1991) and Albera et al. (2004). Finally, Goyal et al. (2013) provides an algorithm with rigorous polynomial time and sampling bounds for underdetermined ICA in high dimension in the presence of Gaussian noise.

Nevertheless, our analysis of the mixture models can be embedded in ICA to show exponential information-theoretic hardness of performing ICA in low-dimension, and thus establishing the blessing of dimensionality for ICA as well.

Theorem 4 *Let X be a set of $4k^2$ uniformly random vectors from $S^{n-1} \subset \mathbb{R}^n$. Then, with high probability, there exist non-empty two disjoint subsets of X such that when these two sets form the columns of matrices A and B respectively, there exist noisy ICA models $AS + \eta$ and $BS' + \eta'$ exponentially close as a function of $(k/\log k)^{\frac{1}{n}}$ as distributions in L^1 distance satisfying: (1) The coordinates of S (and similarly for S') are scaled Poisson distributions, $S_i \sim \alpha_i \mathsf{Poisson}(\lambda_i)$ where each $\lambda_i \leq k^2$ and $\alpha_i \in (\mathrm{poly}(k^{-1}),\ 1)$. At least one λ_i is inverse polynomially (with respect to k) bounded away from 0. (2) The directional variances of η and η' are bounded by $\mathrm{poly}(k)$.*

We sketch the proof of Theorem 4 in Appendix G. Like Theorem 3, choosing X at random ensures that the columns of A and B are not too close. Condition (1) ensures that at least one of the signals S_i and its cumulants are not too close to 0.

Discussion. Most problems become harder in high dimension, often exponentially harder, a behavior known as "the curse of dimensionality." Showing that a complex problem does not become exponentially harder often constitutes major progress in its understanding. In this work we demonstrate a reversal of this curse, showing that the lower dimensional instances are exponentially harder than those in high dimension. This seems to be a rare situation in statistical inference and computation. In particular, while high-dimensional concentration of mass can sometimes be a blessing of dimensionality, in our case the generic computational efficiency of our problem comes from anti-concentration.

We hope that this work will enable better understanding of this unusual phenomenon and its applicability to a wider class of computational and statistical problems.

2. Preliminaries

The singular values of a matrix $A \in \mathbb{R}^{m \times n}$ will be ordered in the decreasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$. By $\sigma_{\min}(A)$ we mean $\sigma_{\min(m,n)}$.

For a real-valued random variable X, the *cumulants* of X are certain polynomials in the moments of X. For $j \geq 1$, the jth cumulant is denoted $\mathrm{cum}_j(X)$. Denoting $\mathrm{m}_j := \mathbb{E}\left(X^j\right)$, we have, for example: $\mathrm{cum}_1(X) = \mathrm{m}_1$, $\mathrm{cum}_2(X) = \mathrm{m}_2 - \mathrm{m}_1^2$, and $\mathrm{cum}_3(X) = \mathrm{m}_3 - 3\mathrm{m}_2\mathrm{m}_1 + 2\mathrm{m}_1^3$. In general, cumulants can be defined as certain coefficients of a Taylor expansion of the logarithm of the moment generating function of X: $\log(\mathbb{E}_X(e^{tX})) = \sum_{j=1}^{\infty} \mathrm{cum}_j(X)\frac{t^j}{j!}$. The first two cumulants are the same as the expectation and the variance, resp. Cumulants have the property that for two independent random variables X, Y we have $\mathrm{cum}_j(X+Y) = \mathrm{cum}_j(X) + \mathrm{cum}_j(Y)$ (assuming that the first j moments exist for both X and Y). Cumulants are degree-j homogeneous, i.e. if $\alpha \in \mathbb{R}$ and X is a random variable, then $\mathrm{cum}_j(\alpha X) = \alpha^j \mathrm{cum}_j(X)$. The third and higher cumulants of the Gaussian distribution are 0.

Gaussian Mixture Model. For $i = 1, 2, \ldots, m$, define Gaussian random vectors $\eta_i \in \mathbb{R}^n$ with distribution $\eta_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ where $\mu_i \in \mathbb{R}^n$ and $\Sigma_i \in \mathbb{R}^{n \times n}$. Let h be an integer-valued random variable which takes on value $i \in [m]$ with probability $w_i > 0$, henceforth called weights. (Hence

$\sum_{i=1}^m w_i = 1$.) Then, the random vector drawn as $Z = \eta_h$ is said to be a Gaussian Mixture Model (GMM) $w_1\mathcal{N}(\mu_1, \Sigma_1) + \ldots + w_m\mathcal{N}(\mu_m, \Sigma_m)$. The sampling of Z can be interpreted as first picking one of the components $i \in [m]$ according to the weights, and then sampling a Gaussian vector from component i. We will be primarily interested in the mixture of identical Gaussians of known covariance. In particular, there exists known $\Sigma \in \mathbb{R}^{n\times n}$ such that $\Sigma_i = \Sigma$ for each i. Letting $\eta \sim \mathcal{N}(0, \Sigma)$, and denoting by $\mathbf{e}_h$ the random variable which takes on the i^{th} canonical vector $\mathbf{e}_i$ with probability w_i, we can write the GMM model as follows:

$$Z = [\mu_1|\mu_2|\cdots|\mu_m]\mathbf{e}_h + \eta\,. \tag{1}$$

In this formulation, $\mathbf{e}_h$ acts as a selector of a Gaussian mean. Conditioning on $h = i$, we have $Z \sim \mathcal{N}(\mu_i, \Sigma)$, which is consistent with the GMM model.

Given samples from the GMM, the goal is to recover the unknown parameters of the GMM, namely the means $\mu_1, \ldots, \mu_m$ and the weights $w_1, \ldots, w_m$.

Underdetermined ICA. In the basic formulation of ICA, the observed random variable $X \in \mathbb{R}^n$ is drawn according to the model $X = AS$, where $S \in \mathbb{R}^m$ is a latent random vector whose components S_i are independent random variables, and $A \in \mathbb{R}^{n\times m}$ is an unknown *mixing matrix*. The probability distributions of the S_i are unknown except that they are not Gaussian. The ICA problem is to recover A to the extent possible. The underdetermined ICA problem corresponds the case $m \geq n$. We cannot hope to recover A fully because if we flip the sign of the i^{th} column of A, or scale this column by some nonzero factor, then the resulting mixing matrix with an appropriately scaled S_i will again generate the same distribution on X as before. There is an additional ambiguity that arises from not having an ordering on the coordinates S_i: If P is a permutation matrix, then PS gives a new random vector with independent reordered coordinates, AP^T gives a new mixing matrix with reordered columns, and $X = AP^TPS$ provides the same samples as $X = AS$ since P^T is the inverse of P. As AP^T is a permutation of the columns of A, this ambiguity implies that we cannot recover the order of the columns of A. However, it turns out that under certain genericity requirements, we can recover A up to these necessary ambiguities, that is to say we can recover the directions (up to sign) of the columns of A, even in the underdetermined setting.

In this paper, it will be important for us to work with an ICA model where there is Gaussian noise in the data: $X = AS + \eta$, where $\eta \sim \mathcal{N}(0, \Sigma)$ is an additive Gaussian noise independent of S, and the covariance of η given by $\Sigma \in \mathbb{R}^{n\times n}$ is in general unknown and not necessarily spherical. We will refer to this model as the noisy ICA model.

We define the flattening operation $\mathrm{vec}\,(\cdot)$ from a tensor to a vector in the natural way. Namely, when $T \in \mathbb{R}^{n^\ell}$ is a tensor, then $\mathrm{vec}\,(T)_{\delta(i_1,\ldots,i_\ell)} = T_{i_1,\ldots,i_\ell}$ where $\delta(i_1, \ldots, i_\ell) = 1 + \sum_{j=1}^{\ell} n^{\ell-j}(i_j - 1)$ is a bijection with indices i_j running from 1 to n. Roughly speaking, each index is being converted into a digit in a base n number up to the final offset by 1. This is the same flattening that occurs to go from a tensor outer product of vectors to the Kronecker product of vectors.

The ICA algorithm from Goyal et al. (2013) to which we will be reducing learning a GMM relies on the shared tensor structure of the derivatives of the second characteristic function and the higher order multi-variate cumulants. This tensor structure motivates the following form of the Khatri-Rao product:

Definition 5 *Given matrices $A \in \mathbb{R}^{n_1\times m}, B \in \mathbb{R}^{n_2\times m}$, a column-wise Khatri-Rao product is defined by $A \odot B := [\mathrm{vec}\,(A_1 \otimes B_1)\,|\cdots|\mathrm{vec}\,(A_m \otimes B_m)]$, where A_i is the i^{th} column of A, B_i is the i^{th} column of B, $\otimes$ denotes the Kronecker product and $\mathrm{vec}\,(A_1 \otimes B_1)$ is flattening of the tensor $A_1 \otimes B_1$ into a vector. The related Khatri-Rao power is defined by $A^{\odot\ell} = A \odot \cdots \odot A$ (ℓ times).*

This form of the Khatri-Rao product arises when performing a change of coordinates under the ICA model using either higher order cumulants or higher order derivative tensors of the second characteristic function.

ICA Results. Theorem 23 (Appendix H.1, from Goyal et al. (2013)) allows us to recover A up to the necessary ambiguities in the noisy ICA setting. The theorem establishes guarantees for an algorithm from Goyal et al. (2013) for noisy underdetermined ICA, **UnderdeterminedICA**. This algorithm takes as input a tensor order parameter d, number of signals m, access to samples according to the noisy underdetermined ICA model with unknown noise, accuracy parameter ϵ, confidence parameter δ, bounds on moments and cumulants M and Δ, a bound on the conditioning parameter σ_m, and a bound on the cumulant order k. It returns approximations to the columns of A up to sign and permutation.

3. Learning GMM means using underdetermined ICA: The basic idea

In this section we give an informal outline of the proof of our main result, namely learning the means of the components in GMMs via reduction to the underdetermined ICA problem. Our reduction will be discussed in two parts. The first part gives the main idea of the reduction and will demonstrate how to recover the means μ_i up to their norms and signs, i.e. we will get $\pm\mu_i/\left\|\mu_i\right\|$. We will then present the reduction in full. It combines the basic reduction with some preprocessing of the data to recover the μ_i's themselves. The reduction relies on some well-known properties of the Poisson distribution stated in the lemma below; its proof can be found in Appendix B.

Lemma 6 *Fix a positive integer k, and let $p_i \geq 0$ be such that $p_1 + \cdots + p_k = 1$. If $X \sim \mathsf{Poisson}(\lambda)$ and $(Y_1, \ldots, Y_k)|_{X=x} \sim \mathrm{Multinom}(x; p_1, \ldots, p_k)$ then $Y_i \sim \mathsf{Poisson}(p_i\lambda)$ for all i and $Y_1, \ldots, Y_k$ are mutually independent.*

Basic Reduction: The main idea. Recall the GMM from equation (1) is $Z = [\mu_1|\cdots|\mu_m]\mathbf{e}_h + \eta$. Henceforth, we will set $A = [\mu_1|\cdots|\mu_m]$. We can write the GMM in the form $Z = A\mathbf{e}_h + \eta$, which is similar in form to the noisy ICA model, except that $\mathbf{e}_h$ does not have independent coordinates. We now describe how a single sample of an approximate noisy ICA problem is generated.

The reduction involves two internal parameters λ and τ that we will set later. We generate a Poisson random variable $R \sim \mathsf{Poisson}(\lambda)$, and we run the following experiment R times: At the i^{th} step, generate sample Z_i from the GMM. Output the sum of the outcomes of these experiments: $Y = Z_1 + \cdots + Z_R$.

Let S_i be the random variable denoting the number of times samples were taken from the i^{th} Gaussian component in the above experiment. Thus, $S_1 + \cdots + S_m = R$. Note that $S_1, \ldots, S_m$ are not observable although we know their sum. By Lemma 6, each S_i has distribution $\mathsf{Poisson}(w_i\lambda)$, and the random variables S_i are mutually independent. Let $S := (S_1, \ldots, S_m)^T$.

For a non-negative integer t, we define $\eta(t) := \sum_{i=1}^{t} \eta_i$ where the η_i are iid according to $\eta_i \sim \mathcal{N}(0, \Sigma)$. In this definition, t can be a random variable, in which case the η_i are sampled independent of t. Using $\sim$ to indicate that two random variables have the same distribution, then $Y \sim AS + \eta(R)$. If there were no Gaussian noise in the GMM (i.e. if we were sampling from a discrete set of points) then the model becomes simply $Y = AS$, which is the ICA model without noise, and so we could recover A up to necessary ambiguities. However, the model $Y \sim AS + \eta(R)$ fails to satisfy even the assumptions of the noisy ICA model, both because $\eta(R)$ is not independent of S and because $\eta(R)$ is not distributed as a Gaussian random vector.

As the covariance of the additive Gaussian noise is known, we may add additional noise to the samples of Y to obtain a good approximation of the noisy ICA model. Parameter τ, the second parameter of the reduction, is chosen so that with high probability we have $R \leq \tau$. Conditioning on the event $R \leq \tau$ we draw X according to the rule $X = Y + \eta(\tau - R) \sim AS + \eta(R) + \eta(\tau - R)$, where $\eta(R)$, $\eta(\tau - R)$, and S are drawn independently conditioned on R. Then, conditioned on $R \leq \tau$, we have $X \sim AS + \eta(\tau)$.

Note that we have only created an approximation to the ICA model. In particular, restricting $\sum_{i=1}^{m} S_i = R \leq \tau$ can be accomplished using rejection sampling, but the coordinate random variables $S_1, \ldots, S_m$ would no longer be independent. We have two models of interest: a noisy ICA model with no restriction on $R = \sum_{i=1}^{m} S_i$ given by

$$X \sim AS + \eta(\tau) \tag{2}$$

and the restricted model

$$X \sim (AS + \eta(\tau))|_{R \leq \tau} \, . \tag{3}$$

We are unable to produce samples from model (2), but it meets the assumptions of the noisy ICA problem. Pretending we have samples from model (2), we can apply Theorem 23 (Appendix (H.1)) to recover the Gaussian means up to sign and scaling. On the other hand, we can produce samples from model (3), and depending on the choice of τ, the statistical distance between models (2) and (3) can be made arbitrarily close to zero. It will be demonstrated that given an appropriate choice of τ, running **UnderdeterminedICA** on samples from model (3) is equivalent to running **UnderdeterminedICA** on samples from model (2) with high probability, allowing for recovery of the Gaussian mean directions $\pm\mu_i / \|\mu_i\|$ up to some error.

Full reduction. To be able to recover the μ_i without sign or scaling ambiguities, we add an extra coordinate to the GMM as follows. The new means μ_i' are μ_i with an additional coordinate whose value is 1 for all i, i.e. $\mu_i' := \left(\mu_i^T, 1\right)^T$. Moreover, this coordinate has no noise. In other words, each Gaussian component now has an $(n+1) \times (n+1)$ covariance matrix $\Sigma' := \left(\begin{smallmatrix} \Sigma & 0 \\ 0 & 0 \end{smallmatrix}\right)$. It is easy to construct samples from this new GMM given samples from the original: If the original samples were $u_1, u_2 \ldots$, then the new samples are $u_1', u_2' \ldots$ where $u_i' := \left(u_i^T, 1\right)^T$. The reduction proceeds similarly to the above on the new inputs.

Unlike before, we will define the ICA mixing matrix to be $A' := \left[\mu_1'/\|\mu_1'\| \middle| \cdots \middle| \mu_m'/\|\mu_m'\|\right]$ such that it has unit norm columns. The role of matrix A in the basic reduction will now be played by A'. Since we are normalizing the columns of A', we have to scale the ICA signal S obtained in the basic reduction to compensate for this: Define $S_i' := \|\mu_i'\| \, S_i$. Thus, the ICA models obtained in the full reduction are:

$$X' = A'S' + \eta'(\tau) \, , \tag{4}$$

$$X' = (A'S' + \eta'(\tau))|_{R \leq \tau} \, , \tag{5}$$

where we define $\eta'(\tau) = \left(\eta(\tau)^T, 0\right)^T$. As before, we have an ideal noisy ICA model (4) from which we cannot sample, and an approximate noisy ICA model (5) which can be made arbitrarily close to (4) in statistical distance by choosing τ appropriately. With appropriate application of Theorem 23 to these models, we can recover estimates (up to sign) $\{\tilde{A}_1', \ldots, \tilde{A}_m'\}$ of the columns of A'.

By construction, the last coordinate of each $\tilde{A}_i'$ now tells us both the sign and magnitude of each μ_i: Let $\tilde{A}_i'(1:n) \in \mathbb{R}^n$ be the vector consisting of the first n coordinates of $\tilde{A}_i'$, and let $\tilde{A}_i'(n+1)$ be the last coordinate of $\tilde{A}_i'$. Then $\mu_i = A_i'(1:n)/A_i'(n+1) \approx \tilde{A}_i'(1:n)/\tilde{A}_i'(n+1)$, with the sign indeterminacy canceling in the division.

Subroutine 1 Single sample reduction from GMM to approximate ICA

Input: Covariance parameter Σ, access to samples from a mixture of m identical Gaussians in $\mathbb{R}^n$ with variance Σ, Poisson threshold τ, Poisson parameter λ,

Output: Y (a sample from model (5)).

Generate R according to $\mathsf{Poisson}(\lambda)$.
If $R > \tau$ return failure.
Let $Y = 0$.
for $j = 1$ to R **do**
  Get a sample Z_j from the GMM.
  Let $Z'_j = (Z_j^T, 1)^T$ to embed the sample in $\mathbb{R}^{n+1}$.
  $Y = Y + Z'_j$.
end for
Let $\Sigma' = \left(\begin{smallmatrix} \Sigma & 0 \\ 0 & 0 \end{smallmatrix}\right)$ (add a row and column of all zeros)
Generate η' according to $\mathcal{N}(0, (\tau - R)\Sigma')$.
$Y = Y + \eta'$.
return Y.

Algorithm 2 Use ICA to learn the means of a GMM

Input: Covariance matrix Σ, number of components m, upper bound on tensor order parameter d, access to samples from a mixture of m identical, spherical Gaussians in $\mathbb{R}^n$ with covariance Σ, confidence parameter δ, accuracy parameter ϵ, upper bound $w \geq \max_i(w_i)/\min_i(w_i)$, upper bound on the norm of the mixture means u, $r \geq (\max_i \|\mu_i\| + 1)/(\min_i \|\mu_i\|)$, and lower bound b so $0 < b \leq \sigma_m(A^{\odot d/2})$.

Output: $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_m\} \subseteq \mathbb{R}^n$ (approximations to the means of the GMM).

Let $\delta_2 = \delta_1 = \delta/2$.
Let $\sigma = \sup_{v \in S^{n-1}} \sqrt{\mathrm{Var}(v^T \eta(1))}$, for $\eta(1) \sim \mathcal{N}(0, \Sigma)$.
Let $\lambda = m$ be the parameter to be used to generate the Poisson random variable in Subroutine 1.
Let $\tau = 4\big(\log(1/\delta_2) + \log(q(\Theta))\big) \max\big((e\lambda)^2, 4Cd^2\big)$ (the threshold used to add noise in the samples from Subroutine 1, C is a universal constant, and $q(\Theta)$ is a polynomial defined as (19) in the proof of Theorem 1).
Let $\epsilon^* = \epsilon\big(\sqrt{1+u^2} + 2(1+u^2)\big)^{-1}$.
Let $M = \max\big((\tau\sigma)^{d+1}, (w/(\sqrt{1+u^2})^{d+1}\big)(d+1)^{d+1}$.
Let $k = d + 1$.
Let $\Delta = w$.
Invoke **UnderdeterminedICA** with access to Subroutine 1, parameters δ_1, ϵ^*, Δ, M, and k to obtain $\tilde{A}'$ (whose columns approximate the normalized means up to sign and permutation). If any calls to Subroutine 1 result in failure, the algorithm will halt completely.
Divide each column of $\tilde{A}'$ by the value of its last entry.
Remove the last row of $\tilde{A}'$ to obtain $\tilde{B}$.
return the columns of $\tilde{B}$ as $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_m\}$.

4. Correctness of the Algorithm and Reduction

Subroutine 1 captures the sampling process of the reduction: Let Σ be the covariance matrix of the GMM, λ be an integer chosen as input, and a threshold value τ also computed elsewhere and

provided as input. Let $R \sim \mathsf{Poisson}(\lambda)$. If R is larger than τ, the subroutine returns a failure notice and the calling algorithm halts immediately. A requirement, then, should be that the threshold is chosen so that the chance of failure is very small; in our case, τ is chosen so that the chance of failure is half of the confidence parameter given to Algorithm 2. The subroutine then goes through the process described in the full reduction: sampling from the GMM, lifting the sample by appending a 1, then adding a lifted Gaussian so that the total noise has distribution $\mathcal{N}(0, \tau\Sigma)$. The resulting sample is from the model given by (5).

Algorithm 2 works as follows: it takes as input the parameters of the GMM (covariance matrix, number of means), tensor order (as required by **UnderdeterminedICA**), error parameters, and bounds on certain properties of the weights and means. The algorithm then calculates various internal parameters: a bound on directional covariances, Poisson parameter λ, threshold parameter τ, error parameters to be split between the "Poissonization" process and the call to **UnderdeterminedICA**, and values explicitly needed by Goyal et al. (2013) for the analysis of **UnderdeterminedICA**. Other internal values needed by the algorithm are denoted by the constant C and polynomial $q(\Theta)$; their values are determined by the proof of Theorem 1. Briefly, C is a constant so that one can cleanly compute a value of τ that will involve a polynomial, called $q(\Theta)$, of all the other parameters. The algorithm then calls **UnderdeterminedICA**, but instead of giving samples from the GMM, it allows access to Subroutine 1. It is then up to **UnderdeterminedICA** to generate samples as needed (bounded by the polynomial in Theorem 1). In the case that Subroutine 1 returns a failure, the entire algorithm process halts, and returns nothing. If no failure occurs, the matrix returned by **UnderdeterminedICA** will be the matrix of normalized means embedded in $\mathbb{R}^{n+1}$, and the algorithm de-normalizes, removes the last row, and then has approximations to the means of of the GMM.

The bounds are used instead of actual values to allow flexibility—in the context under which the algorithm is invoked—on what the algorithm needs to succeed. However, the closer the bounds are to the actual values, the more efficient the algorithm will be.

Sketch of the correctness argument. The proof of correctness of Algorithm 2 has two main parts. For brevity, the details can be found in Appendix A. In the first part, we analyze the sample complexity of recovering the Gaussian means using **UnderdeterminedICA** when samples are taken from the ideal noisy ICA model (4).

In the second part, we note that we do not have access to the ideal model (4), and that we can only sample from the approximate noisy ICA model (5) using the full reduction. Choosing τ appropriately, we use total variation distance to argue that with high probability, running **UnderdeterminedICA** with samples from the approximate noisy ICA model will produce equally valid results as running **UnderdeterminedICA** with samples from the ideal noisy ICA model. The total variation distance bound is explored in section A.2.

These ideas are combined in section A.3 to prove the correctness of Algorithm 2. One additional technicality arises from the implementation of Algorithm 2. Samples can be drawn from the noisy ICA model $X' = (AS' + \eta'(\tau))|_{R \leq \tau}$ using rejection sampling on R. In order to guarantee Algorithm 2 executes in polynomial time, when a sample of R needs to be rejected, Algorithm 2 terminates in explicit failure. To complete the proof, we argue that with high probability, Algorithm 2 does not explicitly fail.

5. Smoothed Analysis

We start with a base matrix $M \in \mathbb{R}^{n \times \binom{n}{2}}$ and add a perturbation matrix $N \in \mathbb{R}^{n \times \binom{n}{2}}$ with each entry coming iid from $\mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. (We restrict the discussion to the second power for simplicity; extension to higher power is straightforward.) As in Goyal et al. (2013), it will be convenient to work with the multilinear part of the Khatri–Rao product: For a column vector $A_k \in \mathbb{R}^n$ define $A_k^{\ominus 2} \in \mathbb{R}^{\binom{n}{2}}$, a subvector of $A_k^{\odot 2} \in \mathbb{R}^{n^2}$, given by $(A_k^{\ominus 2})_{ij} := (A_k)_i (A_k)_j$ for $1 \le i < j \le n$. Then for a matrix $A = [A_1, \ldots, A_m]$ we have $A^{\ominus 2} := [A_1^{\ominus 2}, \ldots, A_m^{\ominus 2}]$.

Theorem 7 *With the above notation, for any base matrix M with dimensions as above, we have, for some absolute constant C,* $\mathsf{Pr}\left(\sigma_{\min}((M+N)^{\ominus 2}) \le \sigma^2/n^7\right) \le 2C/n$.

Theorem 2 follows immediately from the theorem above by noting that $\sigma_{\min}(A^{\odot 2}) \ge \sigma_{\min}(A^{\ominus 2})$.

Proof In the following, for a vector space V (over the reals) $\mathrm{dist}(v, V')$ denotes the distance between vector $v \in V$ and subspace $V' \subseteq V$; more precisely, $\mathrm{dist}(v, V') := \min_{v' \in V'} \|v - v'\|_2$. We will use a lower bound on $\sigma_{\min}(A)$, found in Appendix H.2.

With probability 1, the columns of the matrix $(M+N)^{\ominus 2}$ are linearly independent. This can be proved along the lines of a similar result in Goyal et al. (2013). Fix $k \in \binom{n}{2}$ and let $u \in \mathbb{R}^{\binom{n}{2}}$ be a unit vector orthogonal to the subspace spanned by the columns of $(M+N)^{\ominus 2}$ other than column k. Vector u is well-defined with probability 1. Then the distance of the k'th column C_k from the span of the rest of the columns is given by

$$\begin{aligned} u^T C_k = u^T (M_k + N_k)^{\ominus 2} &= \sum_{1 \le i < j \le n} u_{ij} (M_{ik} + N_{ik})(M_{jk} + N_{jk}) \\ &= \sum_{1 \le i < j \le n} u_{ij} M_{ik} M_{jk} + \sum_{1 \le i < j \le n} u_{ij} M_{ik} N_{jk} + \sum_{1 \le i < j \le n} u_{ij} N_{ik} M_{jk} + \sum_{1 \le i < j \le n} u_{ij} N_{ik} N_{jk} \\ &=: P(N_{1k}, \ldots, N_{nk}). \end{aligned} \tag{6}$$

Now note that this is a quadratic polynomial in the random variables N_{ik}. We will apply the anticoncentration inequality of Carbery and Wright (2001) to this polynomial to conclude that the distance between the k'th column of $(M+N)^{\ominus 2}$ and the span of the rest of the columns is unlikely to be very small (see Appendix H.3 for the precise result).

Using $\|u\|_2 = 1$, the variance of our polynomial in (6) becomes

$$\mathrm{Var}\left(P(N_{1k}, \ldots, N_{nk})\right) = \sigma^2 \left(\sum_j \Big(\sum_{i : i < j} u_{ij} M_{ik} \Big)^2 + \sum_i \Big(\sum_{j : i < j} u_{ij} M_{jk} \Big)^2 \right) + \sigma^4 \sum_{i<j} u_{ij}^2 \ge \sigma^4.$$

In our application, random variables N_{ik} for $i \in [n]$ are not standard Gaussians but are iid Gaussian with variance σ^2, and our polynomial does not have unit variance. After adjusting for these differences using the estimate on the variance of P above, Lemma 25 gives $\mathsf{Pr}\left(|P(N_{1k}, \ldots, N_{nk}) - t| \le \epsilon\right) \le 2C\sqrt{\epsilon/\sigma^2} = 2C\sqrt{\epsilon}/\sigma$. Therefore, $\mathsf{Pr}\left(\exists k : \mathrm{dist}(C_k, C_{-k}) \le \epsilon\right) \le \binom{n}{2} 2C\sqrt{\epsilon}/\sigma$ by the union bound over the choice of k.

Now choosing $\epsilon = \sigma^2/n^6$, Lemma 24 gives $\mathsf{Pr}\left(\sigma_{\min}((M+N)^{\ominus 2}) \le \sigma^2/n^7\right) \le 2C/n$. ∎

We note that while the above discussion is restricted to Gaussian perturbation, the same technique would work for a much larger class of perturbations. To this end, we would require a version of the Carbery-Wright anticoncentration inequality which is applicable in more general situations. We omit such generalizations here.

6. The curse of low dimensionality for Gaussian mixtures

In this section we prove Theorem 3, which informally says that for small n there is a large class of superpolynomially close mixtures in $\mathbb{R}^n$ with fixed variance. This goes beyond the specific example of exponential closeness given in Moitra and Valiant (2010) as we demonstrate that such mixtures are ubiquitous as long as there is no lower bound on the separation between the components.

Specifically, let S be the cube $[0,1]^n \subset \mathbb{R}^n$. We will show that for any two sets of k points X and Y in S, with fill h (we say that X has fill h, if there is a point of X within distance h of any point of S), there exist two mixtures p, q with means on disjoint subsets of $X \cup Y$, which are exponentially close in $1/h$ in the $L^1(\mathbb{R}^n)$ norm. Note that the fill of a sample from the uniform distribution on the cube can be bounded (with high probability) by $O(\sqrt{n}(\frac{\log k}{k})^{1/n})$ (see proof of Theorem 3 below).

We start by defining some of the key objects. Let $K(x,z) = (2\pi)^{-n/2} e^{-\|x-y\|^2/2}$ be the unit Gaussian kernel. Let $\mathcal{K}$ be the integral operator corresponding to the convolution with a unit Gaussian: $\mathcal{K}g(z) = \int_{\mathbb{R}^n} K(x,z)g(x)dx$. Let X be any subset of k points in $[0,1]^n$. Let K_X be the kernel matrix corresponding to X, $(K_X)_{ij} = K(x_i, x_j)$. It is known to be positive definite. For a function $f : [0,1]^n \to \mathbb{R}$, the *interpolant* is defined as $f_{X,k}(x) = \sum w_i K(x_i, x)$, where the coefficients w_i are chosen so that $(\forall i) f_{X,k}(x_i) = f(x_i)$. It is easy to see that such interpolant exists and is unique, obtained by solving a linear system involving K_X.

We will need some properties of the Reproducing Kernel Hilbert Space H corresponding to the kernel K (see (Wendland, 2005, Chapter 10) for an introduction). In particular, we need the bound $\|f\|_\infty \le \|f\|_H$ and the reproducing property, $\langle f(\cdot), K(x,\cdot)\rangle_H = f(x), \forall f \in H$. For a function of the form $\sum w_i K(x_i, x)$ we have $\|\sum w_i K(x_i, x)\|_H^2 = \sum w_i w_j K(x_i, x_j)$.

Lemma 8 *Let g be any positive function with L_2 norm 1 supported on $[0,1]^n$ and let $f = \mathcal{K}g$. If X has fill h, then there exists $A > 0$ such that $\|f - f_{X,k}\|_{L^\infty(\mathbb{R}^n)} < \exp(A(\log h)/h)$.*

Proof From Rieger and Zwicknagl (2010), Theorem 6.1 (taking $\lambda = 0$) we have that for some $A > 0$ and h sufficiently small $\|f - f_{X,k}\|_{L^2([0,1]^n)} < \exp(A\frac{\log h}{h})$. Note that the norm is on $[0,1]^n$ while we need to control the norm on $\mathbb{R}^n$. To do that we need a bound on the RKHS norm of $f - f_{X,k}$. This ultimately gives control of the norm over $\mathbb{R}^n$ because there is a canonical isometric embedding of elements of H interpreted as functions over $[0,1]$ into elements of H interpreted as functions over $\mathbb{R}^n$. We first observe that for any $x_i \in X$, $f(x_i) - f_{X,k}(x_i) = 0$. Thus, from the reproducing property of RKHS, $\langle f - f_{X,k}, f_{X,k}\rangle_H = 0$. Using properties of RKHS with respect to the operator $\mathcal{K}$ (see, e.g., Proposition 10.28 of Wendland (2005))

$$\begin{aligned}\|f - f_{X,k}\|_H^2 &= \langle f - f_{X,k}, f - f_{X,k}\rangle_H = \langle f - f_{X,k}, f\rangle_H = \langle f - f_{X,k}, \mathcal{K}g\rangle_H \\ &= \langle f - f_{X,k}, g\rangle_{L_2([0,1]^n)} \le \|f - f_{X,k}\|_{L^2([0,1]^n)} \|g\|_{L^2([0,1]^n)} < \exp(A(\log h)/h).\end{aligned}$$

Thus $\|f - f_{X,k}\|_{L^\infty(\mathbb{R}^n)} \le \|f - f_{X,k}\|_H < \exp(A(\log h)/h)$. ■

Theorem 9 *Let X and Y be any two subsets of $[0,1]^n$ with fill h. Then there exist two Gaussian mixtures p and q (with positive coefficients summing to one, but not necessarily the same number of components), which are centered on two disjoint subsets of $X \cup Y$ and such that for some $B > 0$, $\|p - q\|_{L^1(\mathbb{R}^n)} < \exp(B(\log h)h)$.*

Proof To simplify the notation we assume that $n = 1$. The general case follows verbatim, except that the interval of integration, $[-1/h, 1/h]$, and its complement need to be replaced by the sphere of radius $1/h$ and its complement respectively.

Let $f_{X,k}$ and $f_{Y,k}$ be the interpolants, for some fixed sufficiently smooth (as above, $f = \mathcal{K}g$) positive function f with $\int_{[0,1]} f(x)dx = 1$. Using Lemma 8, we see that $\|f_{X,k} - f_{Y,k}\|_{L^\infty(\mathbb{R})} < 2\exp(A\frac{\log h}{h})$. Functions $f_{X,k}$ and $f_{Y,k}$ are both linear combinations of Gaussians possibly with negative coefficients and so is $f_{X,k} - f_{Y,k}$. By collecting positive and negative coefficients we write

$$f_{X,k} - f_{Y,k} = p_1 - p_2, \tag{7}$$

where, p_1 and p_2 are mixtures with positive coefficients only.

Put $p_1 = \sum_{i\in S_1} \alpha_i K(x_i, x)$, $p_2 = \sum_{i\in S_2} \beta_i K(x_i, x)$, where S_1 and S_2 are disjoint subsets of $X \cup Y$. Now we need to ensure that the coefficients can be normalized to sum to 1.

Let $\alpha = \sum \alpha_i$, $\beta = \sum \beta_i$. From (7) and by integrating over the interval $[0, 1]$, and since f is strictly positive on the interval, it is easy to see that $\alpha, \beta \geq 1$. We have

$$|\alpha - \beta| = \left| \int_{\mathbb{R}} p_1(x) - p_2(x)dx \right| \leq \|p_1 - p_2\|_{L^1(\mathbb{R})}$$

$$\|p_1 - p_2\|_{L^1(\mathbb{R})} \leq \int_{[-1/h,1/h]} \|f_{X,k} - f_{Y,k}\|_{L^\infty(\mathbb{R})}dx + 2(\alpha + \beta) \int_{x\in[1/h,\infty)} K(0, x - 1)dx.$$

Noticing that the first summand is bounded by $\frac{2}{h}\exp(A\frac{\log h}{h})$ and the integral in the second summand is even smaller (in fact, $O(e^{-1/h^2})$) , it follows immediately, that $|1 - \frac{\beta}{\alpha}| < \exp(A'\frac{\log h}{h})$ for some A' and h sufficiently small.

Hence, we have $\left\|\frac{1}{\alpha}p_1 - \frac{1}{\beta}p_2\right\|_{L^1(\mathbb{R})} \leq \left\|\frac{\beta}{\alpha}p_1 - p_2\right\|_{L^1(\mathbb{R})} \leq \left|1 - \frac{\beta}{\alpha}\right|\|p_1\|_{L^1(\mathbb{R})} + \|p_1 - p_2\|_{L^1(\mathbb{R})}$. Collecting exponential inequalities completes the proof. ■

Proof [of Theorem 3] For convenience we will use a set of $4k^2$ points instead of k^2. Clearly it does not affect the exponential rate. By a simple covering set argument (cutting the cube into m^n cubes with size $1/m$) and basic probability (the coupon collector's problem), we see that the fill h of $2nm^n \log m$ points is at most $O(\sqrt{n}/m)$ with probability $1 - o(1)$. Hence, given k points, we have $h = O(\sqrt{n}(\frac{\log k}{k})^{1/n})$. We see that with a smaller probability (but still close to 1 for large k), we can sample k points $4k$ times and still have the same fill on each group of k.

Pairing the sets of k points into $2k$ pairs of sets arbitrarily and applying Theorem 9 (to $k + k$ points) we obtain $2k$ pairs of exponentially close mixtures with at most $2k$ components each. If one of the pairs has the same number of components, we are done. If not, by the pigeon-hole principle for at least two pairs of mixtures $p_1 \approx q_1$ and $p_2 \approx q_2$ the differences of the number of components (an integer number between 0 and $2k - 2$) must coincide. Assume without loss of generality that p_1 has no more components that q_1 and p_2 has no more components than q_2.Taking $p = \frac{1}{2}(p_1 + q_2)$ and $q = \frac{1}{2}(p_2 + q_1)$ completes the proof. ■

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants IIS RI 1117707 and CCF AF 1350870. We thank anonymous referees for useful comments, and in particular, for drawing our attention to Valiant and Valiant (2013).

References

D. Achlioptas and F. McSherry. On spectral learning of mixture of distributions. In *The 18th Annual Conference on Learning Theory*, 2005.

L. Albera, A. Ferreol, P. Comon, and P. Chevalier. Blind Identification of Overcomplete MixturEs of sources (BIOME). *Lin. Algebra Appl.*, 391:1–30, 2004.

Noga Alon and Joel H Spencer. *The probabilistic method*. Wiley, 2004.

S. Arora and R. Kannan. Learning Mixtures of Arbitrary Gaussians. In *33rd ACM Symposium on Theory of Computing*, 2001.

Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders. In *NIPS*, pages 2384–2392, 2012.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112. IEEE Computer Society, 2010. ISBN 978-0-7695-4244-7.

Mikhail Belkin, Luis Rademacher, and James Voss. Blind signal separation in the presence of Gaussian noise. In *JMLR W&CP*, volume 30: COLT, pages 270–287, 2013.

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. *CoRR*, abs/1311.3651v4, 2014.

Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in R^n. *Mathematical Research Letters*, 8:233–248, 2001.

J-F Cardoso. Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 3109–3112. IEEE, 1991.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. In *Radar and Signal Processing, IEE Proceedings F*, volume 140, pages 362–370, 1993.

Pierre Comon and Christian Jutten, editors. *Handbook of Blind Source Separation*. Academic Press, 2010.

A. Dasgupta. *Probability for Statistics and Machine Learning*. Springer, 2011.

S. Dasgupta. Learning Mixture of Gaussians. In *40th Annual Symposium on Foundations of Computer Science*, 1999.

S. Dasgupta and L. Schulman. A Two Round Variant of EM for Gaussian Mixtures. In *16th Conference on Uncertainty in Artificial Intelligence*, 2000.

J. Feldman, R. A. Servedio, and R. O'Donnell. PAC Learning Axis Aligned Mixtures of Gaussians with No Separation Assumption. In *The 19th Annual Conference on Learning Theory*, 2006.

Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA. *CoRR*, http://arxiv.org/abs/1306.5825, 2013.

Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS*, pages 11–20, 2013.

Maurice Kendall, Alan Stuart, and J. Keith Ord. *Kendall's advanced theory of statistics. Vol. 1*. Halsted Press, sixth edition, 1994. Distribution theory.

A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *51st Annual IEEE Symposium on Foundations of Computer Science (FOCS 2010)*, 2010.

Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: Invariance and optimality. *Annals of Math.*, 171:295–341, 2010.

Ole A Nielsen. *An Introduction to Integration Theory and Measure Theory*. Wiley, 1997.

B.C. Rennie and A.J. Dobson. On Stirling numbers of the second kind. *Journal of Combinatorial Theory*, 7(2):116 – 121, 1969. ISSN 0021-9800. doi: http://dx.doi.org/10.1016/S0021-9800(69)80045-1. URL http://www.sciencedirect.com/science/article/pii/S0021980069800451.

Christian Rieger and Barbara Zwicknagl. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Advances in Computational Mathematics*, 32 (1):103–129, 2010.

John Riordan. Moment recurrence relations for binomial, poisson and hypergeometric frequency distributions. *Annals of Mathematical Statistics*, 8:103–111, 1937.

Halsey Lawrence Royden, Patrick Fitzpatrick, and Prentice Hall. *Real analysis*, volume 4. Prentice Hall New York, 1988.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.*, 62(12):1707–1739, 2009.

Paul Valiant and Gregory Valiant. Estimating the unseen: Improved estimators for entropy and other properties. In *NIPS*, pages 2157–2165, 2013.

S. Vempala and G. Wang. A Spectral Algorithm for Learning Mixtures of Distributions. In *43rd Annual Symposium on Foundations of Computer Science*, 2002.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge University Press Cambridge, 2005.

A. Winkelbauer. Moments and Absolute Moments of the Normal Distribution. *ArXiv e-prints*, September 2012.

Appendix A. Theorem 1 Proof Details

A.1. Error Analysis of the Ideal Noisy ICA Model

The proposed full reduction from Section 3 provides us with two models. The first is a noisy ICA model from which we cannot sample:

$$\text{(Ideal ICA)} \qquad X' = A'S' + \eta'(\tau)\,. \tag{8}$$

The second is a model that fails to satisfy the assumption that S' has independent coordinates, but it is a model from which we can sample:

$$\text{(Approximate ICA)} \qquad X' = (A'S' + \eta'(\tau))|_{R \leq \tau} \,. \tag{9}$$

Both models rely on the choice of two parameters, λ and τ. The dependence on τ is explicit in the models. The dependence on λ can be summarized in the unrestricted model as $S_i = \frac{1}{\|\mu_i'\|} S_i' \sim \mathsf{Poisson}(w_i \lambda)$ independently of each other, and $R = \sum_{i=1}^m S_i \sim \mathsf{Poisson}(\lambda)$.

The probability of choosing $R > \tau$ will be seen to be exponentially small in τ. For this reason, running **UnderdeterminedICA** with polynomially many samples from model (8) will with high probability be equivalent to running the ICA Algorithm with samples from model (9). This notion will be made precise later using total variation distance.

For the remainder of this subsection, we proceed as if samples are drawn from the ideal noisy ICA model (8). Thus, to recover the columns of A', it suffices to run **UnderdeterminedICA** on samples of X'. Theorem 23 can be used for this analysis so long as we can obtain the necessary bounds on the cumulants of S', moments of S', and the moments of $\eta'(\tau)$. We define $w_{\min} := \min_i w_i$ and $w_{\max} := \max_i w_i$. Then, the cumulants of S' are bounded by the following lemma:

Lemma 10 *Given $\ell \in \mathbb{Z}^+$, $\mathrm{cum}_\ell(S_i') \geq w_i \lambda$ for each S_i'. In particular, then $\mathrm{cum}_\ell(S_i') \geq w_{\min} \lambda$.*

Proof By construction, $S_i' = \|\mu_i'\| S_i$. By the homogeneity property of univariate cumulants,

$$\mathrm{cum}_\ell(S_i') = \mathrm{cum}_\ell(\|\mu_i'\| S_i) = \|\mu_i'\|^\ell \mathrm{cum}_\ell(S_i)$$

As $\mu_i'(n+1) = 1$, $\|\mu_i'\| \geq 1$. The cumulants of the Poisson distribution are given in Lemma 21. It follows that $\mathrm{cum}_\ell(S_i') \geq \mathrm{cum}_\ell(S_i) = w_i \lambda$. ■

The bounds on the moments of S_i' for each i can be computed using the following lemma:

Lemma 11 *For $\ell \in \mathbb{Z}^+$, we have $\mathbb{E}\left(S_i'^\ell\right) \leq (\|\mu_i'\| w_i \lambda)^\ell \ell^\ell$.*

Proof Let Y denote a random variable drawn from $\mathsf{Poisson}(\alpha)$. It is known (see Riordan (1937)) that

$$\mathbb{E}\left(Y^\ell\right) = \sum_{i=1}^{\ell} \alpha^i \left\{ {\ell \atop i} \right\}$$

where $\left\{ {\ell \atop i} \right\}$ denotes Stirling number of the second kind. Using Lemma 20, it follows that

$$\mathbb{E}\left(Y^\ell\right) \leq \sum_{i=1}^{\ell} \alpha^i \ell^{\ell-1} \leq \ell \alpha^\ell \ell^{\ell-1} = \alpha^\ell \ell^\ell.$$

Since $S_i' = \mu_i' S_i$ where $S_i \sim \mathsf{Poisson}(\lambda w_i)$, it follows that $\mathbb{E}\left(S_i'^\ell\right) = \|\mu_i'\|^\ell \mathbb{E}\left(S_i^\ell\right) \leq \|\mu_i'\|^\ell (w_i \lambda_i)^\ell \ell^\ell$. ■

The absolute moments of Gaussian random variables are well known. For completeness, the bounds are provided in Lemma 22 of Appendix E.

Defining $\sigma = \sup_{v \in S^{n-1}} \sqrt{\mathrm{Var}(v^T \eta'(1))}$; vectors $\mu'_{\max} = \max_i \|\mu_i'\|$, $\mu'_{\min} = \min_i \|\mu_i'\|$, and similarly $\mu_{\max}$ and $\mu_{\min}$ for later; and choosing $\lambda = m$, we can now show a polynomial bound for the error in recovering the columns of A' using **UnderdeterminedICA**.

Theorem 12 (ICA specialized to the ideal case) *Suppose that samples of X' are taken from the unrestricted ICA model (5) choosing parameter $\lambda = m$ and τ a constant. Suppose that **UnderdeterminedICA** is run using these samples. Suppose $\sigma_m(A'^{\odot d/2}) > 0$. Fix $\epsilon \in (0, 1/2)$ and $\delta \in (0, 1/2)$. Then with probability $1 - \delta$, when the number of samples N is:*

$$N \geq \text{poly}\left(n^d, m^{d^2}, (\tau\sigma)^{d^2}, \left\|\mu'_{\max}\right\|^{d^2}, (w_{\max}/w_{\min})^{d^2}, d^{d^2}, 1/\sigma_m(A'^{\odot d/2})^d, 1/\epsilon, 1/\delta\right) \quad (10)$$

*the columns of A' are recovered within error ϵ up to their signs. That is, denoting the columns returned from **UnderdeterminedICA** by $\tilde{A}'_1, \ldots, \tilde{A}'_m$, there exists $\alpha_1, \cdots, \alpha_m, \in \{-1, +1\}$ and a permutation p of $[m]$ such that $\left\|A'_i - \alpha_i \tilde{A}'_{p(i)}\right\| < \epsilon$ for each i.*

Proof Obtaining the sample bound is an exercise of rewriting the parameters associated with the model $X' = A'S' + \eta'(\tau)$ in a way which can be used by Theorem 23. In what follows, where new parameters are introduced without being described, they will correspond to parameters of the same name defined in and used by the statement of Theorem 23.

Parameter d is fixed. We must choose $k_1, \ldots, k_m$ and k such that $d < k_i \leq k$ and $\text{cum}_{k_i}(S'_i)$ is bounded away from 0. It suffices to choose $k_1 = \cdots = k_m = k = d + 1$. By Lemma 10, $\text{cum}_{d+1}(S'_i) \geq w_{\min}\lambda = w_{\min}m$ for each i. As $w_{\max} \geq \frac{1}{m}\sum_{i=1}^m w_i = \frac{1}{m}$, we have that $\text{cum}_{d+1}(S'_i) \geq \frac{w_{\min}}{w_{\max}}$ for each i, giving a somewhat more natural condition number. In the notation of Theorem 23, we have a constant

$$\Delta = \frac{w_{\min}}{w_{\max}} \quad (11)$$

such that $\text{cum}_{d+1}(S'_i) \geq \Delta$ for each i.

Now we consider the upper bound M on the absolute moments of both $\eta'(\tau)$ and on S'_i. As the Poisson distribution takes on non-negative values, it follows that $S'_i = \|\mu'_i\| S_i$ takes on non-negative values. Thus, the moments and absolute moments of S'_i coincide. Using Lemma 11, we have that $\mathbb{E}\left(|S'_i|^{d+1}\right) = \mathbb{E}\left((S'_i)^{d+1}\right) \leq (\|\mu'_i\| w_i\lambda)^{d+1}(d+1)^{d+1}$. Thus, for M to bound the $(d+1)^{\text{th}}$ moment of S'_i, it suffices that $M \geq (\|\mu'_{\max}\| w_{\max}\lambda)^{d+1}(d+1)^{d+1}$. Noting that

$$w_{\max}\lambda = w_{\max}m = \frac{w_{\max}}{1/m} \leq \frac{w_{\max}}{w_{\min}}$$

it suffices that $M \geq (\|\mu'_{\max}\| \frac{w_{\max}}{w_{\min}})^{d+1}(d+1)^{d+1}$, giving a more natural condition number.

Now we bound the absolute moments of the Gaussian distribution. As $d \in 2\mathbb{N}$, it follows that $d+1$ is odd. Given a unit vector $u \in \mathbb{R}^n$, it follows from Lemma 22 that

$$\mathbb{E}\left(\left|\langle u, \eta'(\tau)\rangle\right|^{d+1}\right) = \text{Var}(\langle u, \eta'(\tau)\rangle)^{\frac{(d+1)}{2}} 2^{d/2}\,(d/2)!\frac{1}{\sqrt{\pi}} = \tau^{d+1}\text{Var}(\langle u, \eta'(1)\rangle)^{\frac{(d+1)}{2}} 2^{d/2}\,(d/2)!\frac{1}{\sqrt{\pi}}\,.$$

σ gives a clear upper bound for $\text{Var}(\langle u, \eta'(1)\rangle)^{1/2}$, and $(d+1)^{d+1}$ gives a clear upper bound to $\frac{1}{\sqrt{\pi}}2^{d/2}(d/2)!$. As such, it suffices that $M \geq (\tau\sigma)^{d+1}(d+1)^{d+1}$ in order to guarantee that $M \geq \mathbb{E}\left(|\langle u, \eta'(\tau)\rangle|^{d+1}\right)$. Using the obtained bounds for M from the Poisson and Normal variables, it suffices that M be taken such that

$$M \geq \max\left((\tau\sigma)^{d+1}, (\left\|\mu'_{\max}\right\| \frac{w_{\max}}{w_{\min}})^{d+1}\right)(d+1)^{d+1} \quad (12)$$

to guarantee that M bounds all required order $d+1$ absolute moments.

We can now apply Theorem 23, using the parameter values $k = d+1$, Δ from (11), and M from (12). Then with probability $1-\delta$,

$$N \geq \text{poly}\left(n^{2d+1}, m^{d^2}, (\tau\sigma)^{d^2}, \left\|\mu'_{\max}\right\|^{d^2}, (w_{\max}/w_{\min})^{d^2}, (d+1)^{d^2}, \right. \\ \left. 1/\sigma_m(A'^{\odot d/2})^{d+1}, 1/\epsilon, 1/\delta\right) \tag{13}$$

samples suffice to recover up to sign the columns of A' within ϵ accuracy. More precisely, letting $\tilde{A}'_1, \ldots, \tilde{A}'_m$ give the columns produced by **UnderdeterminedICA**, then there exists parameters $\alpha_1, \ldots, \alpha_m$ such that $\alpha_i \in \{-1, +1\}$ captures the sign indeterminacy, and a permutation p on $[m]$ such that $\left\|A'_i - \tilde{A}'_{p(i)}\right\| < \epsilon$ for each i.

The poly bound in (13) is equivalent to the poly bound in (10). ■

Theorem 12 allows us to recover the columns of A' up to sign. However, what we really want to recover are the means of the original Gaussian mixture model, which are the columns of A. Recalling the correspondence between A' and A laid out in section 3, the Gaussian means $\mu_1, \ldots, \mu_m$ which form the columns of A are related to the columns $\mu'_1, \ldots, \mu'_m$ of A' by the rule $\mu_i = \mu'_i(1:n)/\mu'_i(n+1)$. Using this rule, we can construct estimate the Gaussian means from the estimates of the columns of A'. By propagating the errors from Theorem 12, we arrive at the following result:

Theorem 13 (Recovery of Gaussian means in Ideal Case) *Suppose that **UnderdeterminedICA** is run using samples of X' from the ideal noisy ICA model (8) choosing parameters $\lambda = m$ and τ a constant. Define $B \in \mathbb{R}^{n \times m}$ such that $B_i = A_i / \|A_i\|$. Suppose further that $\sigma_m(B^{\odot d/2}) > 0$. Let $\tilde{A}'_1, \cdots, \tilde{A}'_m$ be the returned estimates of the columns of A' (from model (8)) by **UnderdeterminedICA**. Let $\tilde{\mu}_i = \tilde{A}'_i(1:n)/\tilde{A}'_i(n+1)$ for each i. Fix error parameters $\epsilon \in (0, 1/2)$ and $\delta \in (0, 1/2)$. When at least*

$$N \geq \text{poly}\left(n^d, m^{d^2}, (\tau\sigma)^{d^2}, \|\mu_{\max}\|^{d^2}, \left(\frac{w_{\max}}{w_{\min}}\right)^{d^2}, d^{d^2}, \left(\frac{\|\mu_{\max}\|+1}{\|\mu_{\min}\|}\right)^{d^2}, \frac{1}{\sigma_m(B^{\odot d/2})^d}, \frac{1}{\epsilon}, \frac{1}{\delta}\right) \tag{14}$$

samples are used, then with probability $1-\delta$ there exists a permutation p of $[m]$ such that $\left\|\tilde{\mu}_{p(i)} - \mu_i\right\| < \epsilon$ for each i.

Proof Let $\epsilon^* > 0$ (to be chosen later) give a desired bound on the errors of the columns of A'. Then, from Theorem 12, using

$$N \geq \text{poly}\left(n^d, m^{d^2}, (\tau\sigma)^{d^2}, \left\|\mu'_{\max}\right\|^{d^2}, (w_{\max}/w_{\min})^{d^2}, d^{d^2}, 1/\sigma_m(A'^{\odot d/2})^d, 1/\epsilon^*, 1/\delta\right) \tag{15}$$

samples suffices with probability $1-\delta$ to produce column estimates $\tilde{A}'_1, \ldots, \tilde{A}'_m$ such that for an unknown permutation p and signs $\alpha_1, \ldots, \alpha_m$, $\alpha_{p(1)}\tilde{A}'_{p(1)}, \ldots, \alpha_{p(m)}\tilde{A}'_{p(m)}$ give ϵ^*-close estimates of the columns $A'_1, \ldots, A'_m$ respectively of A'. In order to avoid notational clutter, we will assume without loss of generality that p is the identity map, and hence that $\left\|\alpha_i\tilde{A}'_i - \alpha A'_i\right\| < \epsilon^*$ holds.

This proof proceeds in two steps. First, we replace the dependencies in (15) on parameters from the lifted GMM model generated by the full reduction with dependencies based on the GMM model we are trying to learn. Then, we propagate the error from recovering the columns $\tilde{A}'_i$ to that of recovering $\tilde{\mu}_i$.

Step 1: GMM Dependency Replacements. In the following two claims, we consider alternative lower bounds for N for recovering column estimators $\tilde{A}'_1, \ldots, \tilde{A}'_m$ which are ϵ^*-close up to sign to the columns of A'. In particular, so long as we use at least as many samples of X' as in (15) when calling **UnderdeterminedICA**, then A' will be recovered with the desired precision with probability $1-\delta$.

Claim *The* $\mathrm{poly}(\|\mu'_{\max}\|^{d^2}, d^{d^2})$ *dependence in* (15) *can be replaced by a* $\mathrm{poly}(\|\mu_{\max}\|^{d^2}, d^{d^2})$ *dependence.*

Proof of Claim. By construction, $\mu'_{\max} = \begin{pmatrix} \mu_{\max} \\ 1 \end{pmatrix}$. By the triangle inequality,

$$\left\|\mu'_{\max}\right\|^{d^2} \leq (\|\mu_{\max}\| + 1)^{d^2}$$

where $(\|\mu_{\max}\| + 1)^{d^2}$ is a polynomial q of $\|\mu_{\max}\|$ with coefficients bounded by $(d^2)^{d^2} = d^{2d^2} = \mathrm{poly}(d^{d^2})$. The maximal power of $\|\mu_{\max}\|$ in $q(\|\mu_{\max}\|)$ is d^{d^2}. It follows that $q(\|\mu_{\max}\|) = \mathrm{poly}(\|\mu_{\max}\|^{d^2}, d^{d^2})$. ▲

Claim *The* $\mathrm{poly}(1/\sigma_m(A'^{\odot d/2})^d)$ *in* (15) *can be replaced by a* $\mathrm{poly}((\frac{\|\mu_{\max}\|+1}{\|\mu_{\min}\|})^{d^2}, 1/\sigma_m(B^{\odot d/2})^d)$ *dependence.*

Proof of Claim First define $\underline{A}'$ to be the unnormalized version of A'. That is, $\underline{A}'_i := \mu'_i$. Then, $\underline{A}' = A'\mathrm{diag}\left(\|\mu'_1\|, \ldots, \|\mu'_m\|\right)$ implies $\underline{A}'^{\odot d/2} = A'^{\odot d/2}\mathrm{diag}\left(\|\mu'_1\|^{d/2}, \ldots \|\mu'_m\|^{d/2}\right)$. Thus, $\sigma_m(\underline{A}'^{\odot d/2}) \leq \sigma_m(A'^{\odot d/2}) \|\mu'_{\max}\|^{d/2}$.

Next, we note that $\underline{A}' = \begin{pmatrix} A \\ \mathbf{1} \end{pmatrix}$ where $\mathbf{1}$ is an all ones row vector. It follows that the rows of $A^{\odot d/2}$ are a strict subset of the rows of $\underline{A}'^{\odot d/2}$. Thus,

$$\sigma_m(A^{\odot d/2}) = \inf_{\|u\|=1} \left\|A^{\odot d/2}u\right\| \leq \inf_{\|u\|=1} \left\|\underline{A}'^{\odot d/2}u\right\| = \sigma_m(\underline{A}'^{\odot d/2}) \,.$$

Finally, we note that $B = A\mathrm{diag}\left(\frac{1}{\|\mu_1\|}, \ldots, \frac{1}{\|\mu_m\|}\right)$ and $B^{\odot d/2} = A^{\odot d/2}\mathrm{diag}\left(\frac{1}{\|\mu_1\|^{d/2}}, \ldots, \frac{1}{\|\mu_m\|^{d/2}}\right)$. It follows that $\sigma_m(B^{\odot d/2}) \leq \sigma_m(A^{\odot d/2})\frac{1}{\|\mu_{\min}\|^{d/2}}$. Chaining together inequalities yields:

$$\sigma_m(B^{\odot d/2}) \leq \frac{\|\mu'_{\max}\|^{d/2}}{\|\mu_{\min}\|^{d/2}}\sigma_m(A'^{\odot d/2}) \quad \text{or alternatively} \quad \frac{\|\mu'_{\max}\|^{d/2}}{\|\mu_{\min}\|^{d/2}} \cdot \frac{1}{\sigma_m(B^{\odot d/2})} \geq \frac{1}{\sigma_m(A'^{\odot d/2})} \,.$$

As $\mu'_{\max} = (\mu_{\max}^T\ 1)^T$, the triangle inequality implies $\|\mu'_{\max}\| \leq \|\mu_{\max}\| + 1$. As we require the dependency of at least $N > \mathrm{poly}((1/\sigma_m(A'^{\odot d/2}))^d)$ samples, it suffices to have the replacement dependency of $N > \mathrm{poly}((\frac{\|\mu_{\max}\|+1}{\|\mu_{\min}\|})^{\frac{d}{2}\cdot d}(1/\sigma_m(B^{\odot d/2})^d) = \mathrm{poly}((\frac{\|\mu_{\max}\|+1}{\|\mu_{\min}\|})^{d^2}(1/\sigma_m(B^{\odot d/2})^d)$ samples. ▲

Thus, it is sufficient to call **UnderdeterminedICA** with

$$N \geq \mathrm{poly}\left(n^d, m^{d^2}, (\tau\sigma)^{d^2}, \|\mu_{\max}\|^{d^2}, \left(\frac{w_{\max}}{w_{\min}}\right)^{d^2}, d^{d^2}, \left(\frac{\|\mu_{\max}\|+1}{\|\mu_{\min}\|}\right)^{d^2}, \frac{1}{\sigma_m(B^{\odot d/2})^d}, \frac{1}{\epsilon^*}, \frac{1}{\delta}\right) \tag{16}$$

samples to achieve the desired ϵ^* accuracy on the returned estimates of the columns of A' with probability $1-\delta$.

Step 2: Error propagation. What remains to be shown is that an appropriate choice of ϵ^* enforces $\|\mu_i - \tilde{\mu}_i\| < \epsilon$ by propagating the error.

Recall that $A'_i = \begin{pmatrix} \mu_i \\ 1 \end{pmatrix} \cdot \left\| \begin{pmatrix} \mu_i \\ 1 \end{pmatrix} \right\|^{-1}$, making $A'_i(n+1) = \frac{1}{\sqrt{1+\|\mu_i\|^2}}$. Thus,

$$A'_i(n+1) \geq \frac{1}{\sqrt{1+\|\mu_{\max}\|^2}}\,. \tag{17}$$

We have that:

$$\begin{aligned}
\|\mu_i - \tilde{\mu}_i\| &= \left\| \frac{A'_i(1:n)}{A'_i(n+1)} - \frac{\tilde{A}'_i(1:n)}{\tilde{A}'_i(n+1)} \right\| \\
&= \left\| \frac{A'_i(1:n)}{A'_i(n+1)} - \frac{\alpha_i \tilde{A}'_i(1:n)}{A'_i(n+1)} + \frac{\alpha_i \tilde{A}'_i(1:n)}{A'_i(n+1)} - \frac{\alpha_i \tilde{A}'_i(1:n)}{\alpha_i \tilde{A}'_i(n+1)} \right\| \\
&\leq \frac{\left\| A'_i(1:n) - \alpha_i \tilde{A}'_i(1:n) \right\|}{|A'_i(n+1)|} + \frac{\left\| \tilde{A}'_i(1:n) \right\| \left| \alpha_i \tilde{A}'_i(n+1) - A'_i(n+1) \right|}{\left| A'_i(n+1) \alpha_i \tilde{A}'_i(n+1) \right|} \\
&\leq \epsilon^* \sqrt{1+\|\mu_{\max}\|^2} + \frac{\left| \alpha_i \tilde{A}'_i(n+1) - A'_i(n+1) \right|}{|A'_i(n+1)| \left[|A'_i(n+1)| - \left| \alpha_i \tilde{A}'_i(n+1) - A'_i(n+1) \right| \right]}
\end{aligned}$$

which follows in part by applying (17) for the left summand and noting that $\tilde{A}'_i$ is a unit vector for the right summand, giving the bound $\left\| \tilde{A}'_i(1:n) \right\| \leq 1$. Continuing with the restriction that $\epsilon^* < \frac{1}{2} \frac{1}{\sqrt{1+\|\mu_{\max}\|^2}}$,

$$\begin{aligned}
\|\mu_i - \tilde{\mu}_i\| &\leq \epsilon^* \sqrt{1+\|\mu_{\max}\|^2} + \frac{\epsilon^* \sqrt{1+\|\mu_{\max}\|^2}}{\left[\frac{1}{\sqrt{1+\|\mu_{\max}\|^2}} - \epsilon^* \right]} \\
&\leq \epsilon^* \left(\sqrt{1+\|\mu_{\max}\|^2} + 2(1+\|\mu_{\max}\|^2) \right).
\end{aligned}$$

Then, in order to guarantee that $\|\mu_i - \tilde{\mu}_i\| < \epsilon$, it suffices to choose ϵ^* such that

$$\epsilon^* \left(\sqrt{1+\|\mu_{\max}\|^2} + 2(1+\|\mu_{\max}\|^2) \right) \leq \epsilon,$$

which occurs when

$$\epsilon^* \leq \frac{\epsilon}{\left(\sqrt{1+\|\mu_{\max}\|^2} + 2(1+\|\mu_{\max}\|^2) \right)}\,. \tag{18}$$

As $\epsilon < \frac{1}{2}$, the restriction $\epsilon^* < \frac{1}{2}\sqrt{1+\|\mu_{\max}^2\|}$ holds automatically for the choice of ϵ^* in (18). The sample bound from (16) contains the dependency $N > \text{poly}(\frac{1}{\epsilon^*}, \|\mu_{\max}\|^{d^2})$. Propagating the error gives a replacement dependency of $N > \text{poly}\left(\frac{1}{\epsilon}, \sqrt{1+\|\mu_{\max}\|^2}, \|\mu_{\max}\|^{d^2}\right) = \text{poly}(\frac{1}{\epsilon}, \|\mu_{\max}\|^{d^2})$ as d is non-negative. This propagated dependency is reflected in (14). ■

A.2. Distance of the Sampled Model to the Ideal Model

An important part of the reduction is that the coordinates of S are mutually independent. Without the threshold τ, this is true (c.f. Lemma 6). However, without the threshold, one cannot know how to add more noise so that the total noise on each sample is iid. We show that we can choose the threshold τ large enough that the samples still come from a distribution with arbitrarily small total variation distance to the one with truly independent coordinates.

Lemma 14 *Fix $\delta > 0$. Let $S \sim \mathsf{Poisson}(\lambda)$ for $\lambda \geq \ln \delta$. If $\tau > e\lambda$, $\tau \geq 1$, and $\tau \geq \ln(1/\delta) - \lambda$, then $\mathsf{Pr}(S > \tau) < \delta$.*

Proof By the Chernoff bound (See Theorem A.1.15 in Alon and Spencer (2004)),

$$\mathsf{Pr}(S > \lambda(1+\epsilon)) \leq \left(e^{\epsilon}(1+\epsilon)^{-(1+\epsilon)}\right)^{\lambda}.$$

For any $\tau > \lambda$, letting $\epsilon = \tau/\lambda - 1$, we get

$$\mathsf{Pr}(S > \tau) \leq \frac{e^{-\lambda}(e\lambda)^{\tau}}{\tau^{\tau}}.$$

Let $b = e\lambda$. To get $\mathsf{Pr}(S > \tau) < \delta$, it suffices that $\tau - \tau \log_b \tau \leq \log_b(\delta e^{\lambda})$. Note that

$$\tau(1 - \log_b \tau) = \tau - \tau \log_b \tau = \log_b\left(b^{\tau}(1/\tau)^{\tau}\right).$$

If $\tau - \tau \log_b \tau \leq \log_b\left(\delta e^{\lambda}\right)$, then we have

$$\log_b\left(b^{\tau}(1/\tau)^{\tau}\right) \leq \log_b(\delta e^{\lambda})$$

which then implies it suffices that

$$\frac{b^{\tau}}{\tau^{\tau}} = \frac{(e\lambda)^{\tau}}{\tau^{\tau}} \leq \lambda^{\tau}/\tau^{\tau} \leq (1/e)^{\tau} \leq \delta e^{\lambda}$$

which holds for $\tau \geq \ln\left(\frac{1}{\delta e^{\lambda}}\right) = \ln(1/\delta) - \lambda$, giving the desired result. ■

Lemma 15 *Let $N, \delta > 0$, $N \in \mathbb{N}$, and $T_1, T_2, \ldots, T_N$ be iid with distribution $\mathsf{Poisson}(\lambda)$. If $\tau \geq \ln(N/\delta) - \lambda$ then*

$$\mathsf{Pr}\left(\bigcup_i \{T_i > \tau\}\right) < \delta.$$

Proof By Lemma 14 $\tau \geq \ln(N/\delta) - \lambda$ implies $\mathsf{Pr}\left(T_i > \tau\right) < \delta/N$ for every i. The union bound gives us the desired result. ■

It should now be easy to see that if we choose our threshold τ large enough, our samples can be statistically close (See Appendix F) to ones that would come from the truly independent distribution. This claim is made formal as follows:

Lemma 16 *Fix $\delta > 0$. Let $\tau > 0$. Let F be a Poisson distribution with parameter λ and have corresponding density f. Let G be a discrete distribution with density $g(x) = f(x)/F(\tau)$ when $0 \leq x \leq \tau$ and 0 otherwise. Then $\mathrm{d}_{TV}(F, G) = 1 - F(\tau)$.*

Proof Since we are working with discrete distributions, we can write

$$d_{TV}(F, G) = \frac{1}{2}\sum_{i=0}^{\infty} |f(i) - g(i)|.$$

Then we can compute

$$\mathrm{d}_{TV}(F, G) = \frac{|F(\tau) - 1|}{2F(\tau)}\sum_{i=0}^{\tau} f(i) + \frac{1}{2}\sum_{i=\tau+1}^{\infty} f(i) = \frac{|F(\tau) - 1|}{2} + \frac{1 - F(\tau)}{2} = 1 - F(\tau)\,.$$

■

A.3. Proof of Theorem 1

We now show that after the reduction is applied, we can use the **UnderdeterminedICA** routine given in Goyal et al. (2013) to learn the GMM. Instead of requiring exact values of each parameter, we simply require a bound on each. The algorithm remains polynomial on those bounds, and hence polynomial on the true values.

Proof For consistency with the notation in Goyal et al. (2013), d in the proof below is twice the value of d in the statement of Theorem 1.

The algorithm is provided parameters: Covariance matrix Σ, upper bound on tensor order d, access to samples from a mixture of m identical spherical Gaussians in $\mathbb{R}^n$ with covariance Σ, confidence δ, accuracy ϵ, upper bound $w \geq \max_i(w_i)/\min_i(w_i)$, upper bound on the norm of the mixture means u, lower bound v so $0 < b \leq \sigma_m(A^{\odot d/2})$, and $r \geq \left(\max_i \|\mu_i\| + 1\right)/(\min_i \|\mu_i\|)\big)$.

The algorithm then needs to fix the number of samples N, sampling threshold τ, Poisson parameter λ, and two new errors δ_1 and δ_2 so that $\delta_1 + \delta_2 \leq \delta$. For simplicity, we will take $\delta_1 = \delta_2 = \delta/2$. Then fix $\sigma = \sup_{v \in S^{n-1}} \sqrt{\mathrm{Var}(v^T\eta(1))}$ for $\eta(1) \sim \mathcal{N}(0, \Sigma)$. Recall that B is the matrix whose ith column is $\mu_i/\|\mu_i\|$. Let A' be the matrix whose ith column is $(\mu_i, 1)/\|(\mu_i, 1)\|$.

Step 1 Assume that after drawing samples from Subroutine 1, the signals S_i are mutually independent (as in the "ideal" model given by (4)) and the mean matrix B satisfies $\sigma_m(B^{\odot d/2}) \geq b > 0$. Then by Theorem 13, with probability of error δ_1, the call to **UnderdeterminedICA** in Algorithm 2 recovers the columns of B to within ϵ and up to a permutation using N samples of complexity

$$p\left(\tau^{d^2}, \Theta\right) = \mathrm{poly}\left(n^d, m^{d^2}, (\tau\sigma)^{d^2}, u^{d^2}, w^{d^2}, d^{d^2}, r^{d^2}, 1/b^d, 1/\epsilon, 1/\delta_1\right)$$

where $p(\tau^{d^2}, \Theta)$ is the bound on N promised by Theorem 13 and Θ is all its arguments except the dependence in τ. So then we have that with at least N samples in this "ideal" case, we can recover approximations to the true means in $\mathbb{R}^n$ up to a permutation and within ϵ distance.

Step 2 We need to show that after getting N samples from the reduction, the resulting distribution is still close in total variation to the independent one. We will choose a new $\delta' = \delta_2/(2N)$. Let $R \sim \mathsf{Poisson}(\lambda)$. Given δ', Lemma 16 shows that for $\tau \geq \ln(1/\delta') - \lambda$, with probability $1 - \delta'$, $R \leq \tau$.

Take N iid random variables $X_1, X_2, \ldots, X_N$ from the $\mathsf{Poisson}(\lambda)$ distribution. Let G be a distribution given by density function $g(x) = (f(x)\mathbf{1}_{0\leq x\leq\tau})/F(\tau)$. Let $Y_1, Y_2, \ldots, Y_N$ be iid random variables with distribution G. Denote the joint distribution of the X_i's by F' with density f', and the joint distribution of the Y_i's as G' with density g'. By the union bound and the fact that total variation distance satisfies the triangle inequality,

$$\mathrm{d}_{TV}(F', G') \leq \sum_{i=1}^{N} \mathrm{d}_{TV}(F, G) = N\,\mathrm{d}_{TV}(F, G).$$

Then for our choice of τ, by Lemma 14 and Lemma 16, we have

$$\mathrm{d}_{TV}(F', G') \leq N d_{TV}(F, G) = N\mathsf{Pr}\left(X_1 > \tau\right) \leq N\delta' = \delta_2/2.$$

By the same union bound argument, the probability that the algorithm fails (when $R > \tau$) is at most $\delta_2/2$, since it has to draw N samples. So with high probability, the algorithm does not fail; otherwise, it still does not take more than polynomial time, and will terminate instead of returning a false result.

Step 3 We know that N is at least a polynomial which can be written in terms of the dependence on τ as $p(\tau^{d^2}, \Theta)$. This means there will be a power of τ which dominates all of the τ factors in p, and in particular, will be τ^{Cd^2} for some C. It then suffices to choose C so that $p\left(\tau^{d^2}, \Theta\right) \leq \tau^{Cd^2} q(\Theta) \leq N$, where

$$q(\Theta) = \mathrm{poly}\left(n^d, m^{d^2}, \sigma^{d^2}, u^{d^2}, w^{d^2}, d^{d^2}, r^{d^2}, 1/b^d, 1/\epsilon, 1/\delta_1\right). \tag{19}$$

Then, with the proper choice of τ (to be specified shortly), from step 2 we have

$$p\left(\tau^{d^2}, \Theta\right) \leq \tau^{Cd^2} q(\Theta) \leq N = \frac{\delta_2}{\delta'} \leq \frac{\delta_2 \tau^\tau e^\lambda}{(e\lambda)^\tau} = \frac{\delta \tau^\tau e^\lambda}{2(e\lambda)^\tau}.$$

Since $\lambda \geq 1$ it suffices to choose τ so that

$$\frac{2}{\delta} q(\Theta) \tau^{Cd^2} \leq \frac{\tau^\tau}{\tau^{Cd^2}(e\lambda)^\tau}. \tag{20}$$

Finally, we claim that

$$\tau = 4\big(\log(2/\delta) + \log(q(\Theta))\big) \max\left((e\lambda)^2, 4Cd^2\right) = O\left((\lambda^2 + d^2)\log\frac{q(\Theta)}{\delta}\right)$$

is enough for the desired bound on the sample size. Observe that $4(\log(2/\delta) + \log(q(\Theta))) \geq 1$.

An useful fact is that for general $x, a, b \geq 1$, $x \geq \max(2a, b^2)$ satisfies $x^a \leq x^x/b^x$. This captures the essence of our situation nicely. Letting $e\lambda$ play the role of b, Cd^2 play the role of a and x play the role of τ, to satisfy (20), it suffices that

$$\frac{2}{\delta}q(\Theta) \leq \frac{\tau^{\tau/2}\tau^{\tau/4}\tau^{\tau/4}}{\tau^{Cd^2}(e\lambda)^2}.$$

We can see that $\tau^{\tau/2} \geq (e\lambda)^2$ and $\tau^{\tau/4} \geq \tau^{Cd^2}$ by construction. But we also get $\tau/4 \geq \log(2/\delta) + \log q(\Theta)$ which implies $\tau^{\tau/4} \geq e^{\tau/4} \geq \frac{2}{\delta}q(\Theta)$. Thus for our choice of τ, which also preserves the requirement in Step 2, there is a corresponding set of choices for N, where the required sample size remains polynomial as

$$\text{poly}\left(n^d, m^{d^2}, (\tau\sigma)^{d^2}, u^{d^2}, w^{d^2}, d^{d^2}, r^{d^2}, 1/b^d, 1/\epsilon, 1/\delta\right)$$

where we used the bound $q(\Theta) \leq (n^d m^{d^2}\sigma^{d^2}u^{d^2}w^{d^2}(d+1)^{d^2}r^{d^2}/b^d\delta_1\epsilon)^{O(1)}$. By the choice of τ, one can absorb τ^{d^2} into the above $\text{poly}(\cdot)$ expression, giving the result. ■

Appendix B. Lemmas on the Poisson Distribution

The following lemmas are well-known; see, e.g., Dasgupta (2011). We provide proofs for completeness.

Lemma 17 *If* $X \sim \mathsf{Poisson}(\lambda)$ *and* $Y|_{X=x} \sim \text{Bin}(x, p)$ *then* $Y \sim \mathsf{Poisson}(p\lambda)$.

Proof

$$\begin{aligned}
\Pr(Y = y) &= \sum_{x:x\geq y}^{\infty} \Pr(Y = y \mid X = x)\Pr(X = x) \\
&= \sum_{x:x\geq y}^{\infty} \binom{x}{y} p^y (1-p)^{x-y} \frac{\lambda^x e^{-\lambda}}{x!} \\
&= p^y e^{-\lambda} \sum_{x:x\geq y}^{\infty} \frac{\lambda^x}{x!}\binom{x}{y}(1-p)^{x-y} \\
&= \frac{(p\lambda)^y e^{-\lambda}}{y!} \sum_{x:x\geq y}^{\infty} \frac{(\lambda(1-p))^{x-y}}{(x-y)!} \\
&= \frac{(p\lambda)^y e^{-\lambda}}{y!} e^{(1-p)\lambda} \\
&= \frac{(p\lambda)^y e^{-p\lambda}}{y!}.
\end{aligned}$$

■

Lemma 18 *Fix a positive integer k, and let $p_i \geq 0$ be such that $p_1 + \cdots + p_k = 1$. If $X \sim \mathsf{Poisson}(\lambda)$ and $(Y_1, \ldots, Y_k)|_{X=x} \sim \mathrm{Multinom}(x; p_1, \ldots, p_k)$ then $Y_i \sim \mathsf{Poisson}(p_i\lambda)$ for all i and $Y_1, \ldots, Y_k$ are mutually independent.*

Proof The first part of the lemma (i.e., $Y_i \sim \mathsf{Poisson}(p_i\lambda)$ for all i) follows from Lemma 17. For the second part, let's prove it for the binomial case ($k = 2$); the general case is similar.

$$\begin{aligned}
\mathsf{Pr}\left(Y_1 = y_1, Y_2 = y_2\right) &= \mathsf{Pr}\left(Y_1 = y_1, Y_2 = y_2 \mid X = y_1 + y_2\right) \mathsf{Pr}\left(X = y_1 + y_2\right) \\
&= \binom{y_1 + y_2}{y_1} p^{y_1} (1-p)^{y_2} \cdot \frac{\lambda^{y_1+y_2} e^{-\lambda}}{(y_1+y_2)!} \\
&= \frac{(p\lambda)^{y_1} e^{-p\lambda}}{y_1!} \cdot \frac{((1-p)\lambda)^{y_2} e^{-(1-p)\lambda}}{y_2!} \\
&= \mathsf{Pr}\left(Y_1 = y_1\right) \cdot \mathsf{Pr}\left(Y_2 = y_2\right).
\end{aligned}$$

■

Appendix C. Properties of Cumulants

The following properties of multivariate cumulants are well known and are largely inherited from the definition of the cumulant generating function:

- (Symmetry) Let σ give a permutation of k indices. Then, $\kappa_Y^{i_1,\cdots,i_\ell} = \kappa_Y^{\sigma(i_1),\cdots,\sigma(i_\ell)}$.
- (Multilinearity of coordinate random variables) Given constants $\alpha_1, \cdots, \alpha_\ell$, then

$$\mathrm{cum}(\alpha_1 Y_{i_1}, \cdots, \alpha_\ell Y_{i_\ell}) = \left(\prod_{i=1}^{\ell} \alpha_i\right) \mathrm{cum}(Y_{i_1}, \cdots, Y_{i_\ell}).$$

  Also, given a scalar random variable Z, then

$$\mathrm{cum}(Y_{i_1} + Z, Y_{i_2}, \cdots, Y_{i_\ell}) = \mathrm{cum}(Y_{i_1}, Y_{i_2}, \cdots, Y_{i_\ell}) + \mathrm{cum}(Z, Y_{i_2}, \cdots, Y_{i_\ell})$$

  with symmetry implying the additive multilinear property for all other coordinates.

- (Independence) If there exists i_j, i_k such that Y_{i_j} and Y_{i_k} are independent random variables, then the cross-cumulant $\kappa_Y^{i_1,\cdots,i_\ell} = 0$. Combined with multilinearity, it follows that when there are two independent random vectors Y and Z, then $\kappa_{Y+Z} = \kappa_Y + \kappa_Z$.
- (Vanishing Gaussians) When $\ell \geq 3$, then for the Gaussian random variable η, $\kappa_\eta = 0$.

Appendix D. Bounds on Stirling Numbers of the Second Kind

The following bound comes from (Rennie and Dobson, 1969, Theorem 3).

Lemma 19 *If $n \geq 2$ and $1 \leq r \leq n-1$ are integers, then $\left\{ {n \atop r} \right\} \leq \frac{1}{2}\binom{n}{r} r^{n-r}$.*

From this, we can derive a somewhat looser bound on the Stirling numbers of the second kind which does not depend on r:

Lemma 20 *If $n, r \in \mathbb{Z}^+$ such that $r \leq n$, then $\left\{ {n \atop r} \right\} \leq n^{n-1}$.*

Proof The Stirling number $\left\{ {n \atop k} \right\}$ of the second kind gives a count of the number of ways of splitting a set of n labeled objects into k unlabeled subsets. In the case where $r = n$, then $\left\{ {n \atop r} \right\} = 1$ As $n \geq 1$, it is clear that for these choices of n and r, $\left\{ {n \atop r} \right\} \leq n^{n-1}$. By the restriction $1 \leq r \leq n$, when $n = 1$, then $n = r$ giving that $\left\{ {n \atop r} \right\} = 1$. As such, the only remaining cases to consider are when $n \geq 2$ and $1 \leq r \leq n-1$, the cases where Lemma 19 applies.

When $n \geq 2$ and $1 \leq r \leq n-1$, then

$$\left\{ {n \atop r} \right\} \leq \frac{1}{2}\binom{n}{r} r^{n-r} = \frac{1}{2}\frac{n!}{r!(n-r)!} r^{n-r} \leq \frac{1}{2} n^r r^{n-r-1} < \frac{1}{2} n^r n^{n-r-1} = \frac{1}{2} n^{n-1} ,$$

which is slightly stronger than the desired upper bound. ■

Appendix E. Values of Higher Order Statistics

In this appendix, we gather together some of the explicit values for higher order statistics of the Poisson and Normal distributions required for the analysis of our reduction from learning a Gaussian Mixture Model to learning an ICA model from samples.

Lemma 21 (Cumulants of the Poisson distribution) *Let $X \sim \mathsf{Poisson}(\lambda)$. Then, $\mathrm{cum}_\ell(X) = \lambda$ for every positive integer ℓ.*

Proof The moment generating function of the Poisson distribution is given by $M(t) = \exp(\lambda(e^t - 1))$. The cumulant generating function is thus $g(t) = \log(M(t)) = \lambda(e^t - 1)$. The ℓ^{th} derivative ($\ell \geq 1$) is given by $g^{(\ell)}(t) = \lambda e^t$.

By definition, $\mathrm{cum}_\ell(X) = g^{(\ell)}(0) = \lambda$. ■

Lemma 22 (Absolute moments of the Gaussian distribution) *The absolute moments of the Gaussian random variable $\eta \sim N(0, \sigma^2)$ are given by:*

$$\mathbb{E}\left(|\eta|^\ell\right) = \begin{cases} \sigma^\ell \frac{\ell!}{2^{\ell/2}(\ell/2)!} & \text{if } \ell \text{ is even} \\ \sigma^\ell 2^{\ell/2} (\frac{\ell-1}{2})! \frac{1}{\sqrt{\pi}} & \text{if } \ell \text{ is odd.} \end{cases}$$

The case that ℓ is even in Lemma 22 is well known, and can be found for instance in (Kendall et al., 1994, Section 3.4). For general ℓ, it is known (see Winkelbauer (2012)) that

$$\mathbb{E}\left(|\eta|^\ell\right) = \sigma^\ell 2^{\ell/2} \Gamma\left(\frac{\ell+1}{2}\right) \frac{1}{\sqrt{\pi}} .$$

When ℓ is odd, $\frac{\ell+1}{2}$ is an integer, allowing the Gamma function to simplify to a factorial: $\Gamma\left(\frac{\ell+1}{2}\right) = \left(\frac{\ell-1}{2}\right)!$. This gives the case where ℓ is odd in Lemma 22.

Appendix F. Total Variation Distance

Total variation is a type of statistical distance metric between probability distributions. In words, the total variation between two measures is the largest difference between the measures on a single event. Clearly, this distance is bounded above by 1.

For probability measures F and G on a sample space Ω with sigma-algebra Σ, the total variation is denoted and defined as:

$$\mathrm{d}_{TV}(F, G) := \sup_{A \in \Sigma} |F(A) - G(A)|.$$

Equivalently, when F and G are distribution functions having densities f and g, respectively,

$$\mathrm{d}_{TV}(F, G) = \frac{1}{2} \int_{\Omega} |f - g| d\mu$$

where μ is an arbitrary positive measure for which F and G are absolutely continuous.

More specifically, when F and G are discrete distributions with known densities, we can write

$$\mathrm{d}_{TV}(F, G) = \frac{1}{2} \sum_{k=0}^{\infty} |f(k) - g(k)|$$

where we choose μ that simply assigns unit measure to each atom of Ω (in this case, absolute continuity is trivial since $\mu(A) = 0$ only when A is empty and thus $F(A)$ must also be 0). For more discussion, one can see Definition 15.3 in Nielsen (1997) and Sect. 11.6 in Royden et al. (1988).

Appendix G. Sketch for the proof of Theorem 4

Lower bound for ICA. We can use our Poissonization technique to embed difficult instances of learning GMMs into the ICA setting to prove that ICA is information-theoretically hard when the observed dimension n is a constant using the lower bound for learning GMMs. We are not aware of any existing lower bounds in the literature for this problem. We only provide an informal outline of the argument.

Theorem 3 gives us two GMMs p and q of identity covariance Gaussians that are exponentially close with respect to $(k/\log k)^{\frac{1}{n}}$ (where $4k^2$ points are used to generate the Gaussian means) in L^1 distance but far in parameter distance. We apply the basic reduction from Section 3 with λ set to the number of Gaussian means associated with the respective GMMs p and q to obtain the ideal noisy ICA models $X_p = A_p S_p + \eta(\tau)$ and $X_q = A_q S_q + \eta(\tau)$ (see the construction of model (2) from Section 3). Here, $\eta \sim \mathcal{N}(0, I)$, and the choice of τ will be specified later. Recall that $R_p = \sum_i S_{pi} \sim \mathsf{Poisson}(m_p)$ and $R_q = \sum_i S_{qi} \sim \mathsf{Poisson}(m_q)$ with parameters m_p and m_q denoting the number of columns of A_p and A_q respectively. Let $w_1, \ldots, w_{m_p}$ be the Gaussian weights associated with the GMM p. By Lemma 6, each $S_{pi} \sim \mathsf{Poisson}(w_i m_p)$. As there must exist $w_i \in [\frac{1}{m_p}, 1]$, and as $m_p \in [1, 4k^2]$, it follows that there exists i such that $S_{pi} \sim \mathsf{Poisson}(\lambda)$ for some $\lambda \in [\frac{1}{4k^2}, 4k^2]$. The same result holds for S_q.

Now, we let S_p and S_q take on the scaling information of the ICA model by setting $\alpha_{pi} = \|A_{pi}\|$, $\alpha_{qi} = \|A_{qi}\|$ and replacing S_{pi} and S_{qj} by $\alpha_{pi} S_{pi}$ and $\alpha_{qj} S_{qj}$ respectively, and replacing the columns of A_p and A_q with their unit-normalized versions. While Theorem 3 is proven in the setting where Gaussian means are drawn uniformly at random from the unit hypercube, it can be reformulated to have Gaussian means drawn uniformly at random from the unit ball. Under such a reformulation,

the resulting normalized columns of A_p and A_q are chosen from a set of $4k^2$ unit vectors drawn uniformly from the unit sphere $S^{n-1} \subset \mathbb{R}^n$. Further, with high probability, each α_{pi} and each α_{qi} is inverse polynomially (with respect to k) bounded away from 0.

Lemma 14 implies that for a choice of τ_p which is linear in $m_p \leq 4k^2$, the probability of a draw with $R_p > \tau_p$ is exponentially small with respect to τ_p. For such choices of τ_p and τ_q which are linear in m_p and m_q respectively, we choose $\tau = \max(\tau_p, \tau_q, k)$ as the common threshold for the above ICA models. Note that since τ is at most linear in $4k^2$, the directional variances of $\eta(\tau)$ are also linear in $4k^2$, and hence polynomially bounded in k as desired.

Since the L^1 (or equivalently total variation) distance between p and q is exponentially small with respect to $(k/\log k)^{\frac{1}{n}}$, the total variation distance between the two resulting ICA models—the distributions of X_p and X_q respectively—is also exponentially small with respect to $(k/\log k)^{\frac{1}{n}}$. To see this, we must condition on several cases. First, conditioning either model on $R > \tau$, we have that $\Pr(R > \tau)$ is exponentially small with respect to τ and hence k, and its contribution to the overall total variation distance between X_p and X_q is thus exponentially small. Conditioning on $R = z$ where $z \in \{0, 1, \ldots, \tau\}$, then the facts that p and q are close in total variation distance and that total variation distance satisfies a version of the triangle inequality—for random variables $C,\ D,\ E$, and F, we have $d_{TV}(C+D, E+F) \leq d_{TV}(C, E) + d_{TV}(D, F)$—imply that by viewing X_p (and similarly for X_q) as the sum of z draws from the distribution p and $\tau - z$ draws from the additive Gaussian noise distribution η, the total variation distance between X_p and X_q conditioned on $R = z$ is still exponentially small. Thus, the non-conditional distributions of X_p and X_q will be exponentially close with respect to $(k/\log k)^{\frac{1}{n}}$ in total variation distance. In particular, the sample complexity of distinguishing between X_p and X_q is at least exponential in $(k/\log k)^{\frac{1}{n}}$.

One can also interpret ICA with Gaussian noise as ICA without noise by treating the noise as extra signals: If $X = AS + \eta$ is an ICA model where $A \in \mathbb{R}^{n \times m}$ and $\eta \in \mathbb{R}^n$ is spherical Gaussian noise, then by defining $A' := [A|I_n]$, and $S' := [S^T, \eta^T]^T$ we get $X = A'S'$ which is a noiseless model with some of the signals being Gaussian. In such cases, algorithms (such as that of Goyal et al. (2013)) are able to still recover the non-Gaussian portion A of A'. Our result shows that such algorithms cannot be efficient if the observations are in small dimensions (i.e. n is small).

Appendix H.

H.1. Underdetermined ICA theorem

Theorem 23 (Goyal et al. (2013)) *Let a random vector $x \in \mathbb{R}^n$ be given by an underdetermined ICA model with unknown Gaussian noise $x = As + \eta$ where $A \in \mathbb{R}^{n \times m}$ with $m \geq n$ has unit norm columns, and both A and the covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ are unknown. Let $d \in 2\mathbb{N}$ be such that $\sigma_m(A^{\odot d/2}) > 0$. Let $k > d$ be such that for each s_i, there is a k_i satisfying $d < k_i \leq k$ and $|\mathrm{cum}_{k_i}(s_i)| \geq \Delta$, and $\mathbb{E}\left(|s_i|^k\right) \leq M$. Moreover, suppose that the noise also satisfies the same moment condition: $\mathbb{E}\left(|\langle u, \eta_i \rangle|^k\right) \leq M$ for any unit vector $u \in \mathbb{R}^n$ (this is satisfied if we have $k!\sigma^k \leq M$ where σ^2 is the maximum eigenvalue of Σ). Then algorithm **UnderdeterminedICA** returns a set of n-dimensional vectors $(\tilde{A}_i)_{i=1}^m$ so that for some permutation π of $[m]$ and signs $\alpha_i \in \{-1, 1\}$ we have $\left\|\alpha_i \tilde{A}_{\pi(i)} - A_i\right\| \leq \epsilon$ for all $i \in [m]$. Its sample and time complexity are poly $\left(n^k, m^{k^2}, M^k, 1/\Delta^k, 1/\sigma_m(A^{\odot d/2})^k, 1/\epsilon, 1/\delta\right)$.*

H.2. Rudelson-Vershynin subspace bound

Lemma 24 (Rudelson and Vershynin (2009)) *If $A \in \mathbb{R}^{n\times m}$ has columns $C_1, \ldots, C_m$, then denoting $C_{-i} = \operatorname{span}(C_j : j \neq i)$, we have*

$$\frac{1}{\sqrt{m}} \min_{i\in[m]} \operatorname{dist}(C_i, C_{-i}) \leq \sigma_{\min}(A),$$

where as usual $\sigma_{\min}(A) = \sigma_{\min(m,n)}(A)$.

H.3. Carbery-Wright anticoncentration

The version of the anticoncentration inequality we use is explicitly given in Mossel et al. (2010) which in turn follows immediately from Carbery and Wright (2001):

Lemma 25 (Mossel et al. (2010)) *Let $Q(x_1, \ldots, x_n)$ be a multilinear polynomial of degree d. Suppose that $\operatorname{Var}(Q) = 1$ when $x_i \sim \mathcal{N}(0,1)$ for all i. Then there exists an absolute constant C such that for $t \in \mathbb{R}$ and $\epsilon > 0$,*

$$\Pr_{(x_1,\ldots,x_n)\sim\mathcal{N}(0,I_n)}(|Q(x_1,\ldots,x_n) - t| \leq \epsilon) \leq Cd\epsilon^{1/d}.$$

Appendix I. Recovery of Gaussian Weights

Multivariate cumulant tensors and their properties. Our technique for the recovery of the Gaussian weights relies on the tensor properties of multivariate cumulants that have been used in the ICA literature.

Given a random vector $Y \in \mathbb{R}^n$, the moment generating function of Y is defined as $M_Y(t) := \mathbb{E}_Y(\exp(t^T Y))$. The *cumulant generating function* is the logarithm of the moment generating function: $g_Y(t) := \log(\mathbb{E}_Y(\exp(t^T Y))$.

Similarly to the univariate case, multivariate cumulants are defined using the coefficients of the Taylor expansion of the cumulant generating function. We use both $\kappa_Y^{j_1,\ldots,j_\ell}$ and $\operatorname{cum}(Y_{j_1}, \ldots, Y_{j_\ell})$ to denote the order-ℓ cross cumulant between the random variables $Y_{j_1}, Y_{j_2}, \ldots, Y_{j_\ell}$. Then, the cross-cumulants $\kappa_Y^{j_1,\ldots,j_\ell}$ are given by the formula $\kappa_Y^{j_1,\ldots,j_\ell} = \frac{\partial}{\partial t_{j_1}} \cdots \frac{\partial}{\partial t_{j_\ell}} g_Y(t)\big|_{t=0}$. When unindexed, κ_Y will denote the full order-ℓ tensor containing all cross-cumulants, with the order of the tensor being made clear by context. In the special case where $j_1 = \cdots = j_\ell = j$, we obtain the order-$\ell$ univariate cumulant $\operatorname{cum}_\ell(Y_j) = \kappa_Y^{j,\ldots,j}$ (j repeated ℓ times) previously defined. We will use some well known properties of multivariate cumulants, found in Appendix C.

The most theoretically justified ICA algorithms have relied on the tensor structure of multivariate cumulants, including the early, popular practical algorithm JADE Cardoso and Souloumiac (1993). In the fully determined ICA setting in which the number source signals does not exceed the ambient dimension, the papers Arora et al. (2012) and Belkin et al. (2013) demonstrate that ICA with additive Gaussian noise can be solved in polynomial time and using polynomial samples. The tensor structure of the cumulants was (to the best of our knowledge) first exploited in Cardoso (1991) and later in Albera et al. (2004) to solve underdetermined ICA. Finally, Goyal et al. (2013) provides an algorithm with rigorous polynomial time and sampling bounds for underdetermined ICA in the presence of Gaussian noise.

Weight recovery (main idea). Under the basic ICA reduction (see section 3) using the Poisson distribution with parameter λ, we have that $X = AS + \eta$ is observed such that $A = [\mu_1|\cdots|\mu_m]$ and $S_i \sim \mathsf{Poisson}(w_i\lambda)$. As A has already been recovered, what remains to be recovered are the weights $w_1, \cdots, w_m$. These can be recovered using the tensor structure of higher order cumulants. The critical relationship is captured by the following Lemma:

Lemma 26 *Suppose that $X = AS + \eta$ gives a noisy ICA model. When κ_X is of order $\ell > 2$, then* $\mathrm{vec}\,(\kappa_X) = A^{\odot\ell}(\mathrm{cum}_\ell(S_1), \ldots, \mathrm{cum}_\ell(S_m))^T$.

Proof It is easily seen that the Gaussian component has no effect on the cumulant:

$$\kappa_X = \kappa_{AS+\eta} = \kappa_{AS} + \kappa_\eta = \kappa_{AS}$$

Then, we expand κ_X:

$$\begin{aligned}
\kappa_X^{i_1,\cdots,i_\ell} &= \kappa_{AS}^{i_1,\cdots,i_\ell} = \mathrm{cum}((AS)_{i_1}, \cdots, (AS)_{i_\ell}) \\
&= \mathrm{cum}\left(\sum_{j_1=1}^m A_{i_1 j_1} S_{j_1}, \cdots, \sum_{j_\ell=1}^m A_{i_\ell j_\ell} S_{j_\ell}\right) \\
&= \sum_{j_1,\cdots,j_\ell \in [m]} \left(\prod_{k=1}^\ell A_{i_k j_k}\right) \mathrm{cum}(S_{j_1}, \cdots, S_{j_\ell}) \qquad \text{by multilinearity}
\end{aligned}$$

But, by independence, $\mathrm{cum}(S_{j_1}, \cdots, S_{j_m}) = 0$ whenever $j_1 = j_2 = \cdots = j_\ell$ fails to hold. Thus,

$$\kappa_X^{i_1,\cdots,i_\ell} = \sum_{j=1}^m \left(\prod_{k=1}^\ell A_{i_k j}\right) \mathrm{cum}_\ell(S_j) = \sum_{j=1}^m \left((A_j)^{\otimes\ell}\right)_{i_1,\cdots,i_\ell} \mathrm{cum}_\ell(S_j)$$

Flattening yields: $\mathrm{vec}\,(\kappa_X) = A^{\odot\ell}(\mathrm{cum}_\ell(S_1), \cdots, \mathrm{cum}_\ell(S_m))^T$. ■

In particular, we have that $S_i \sim \mathsf{Poisson}(w_i\lambda)$ with w_i the probability of sampling from the i^{th} Gaussian. Given knowledge of A and the cumulants of the Poisson distribution, we can recover the Gaussian weights.

Theorem 27 *Suppose that $X = AS + \eta(\tau)$ is the unrestricted noisy ICA model from the basic reduction (see section 3). Let $\ell > 2$ be such that $A^{\odot\ell}$ has linearly independent columns, and let $(A^{\odot\ell})^\dagger$ be its Moore-Penrose pseudoinverse. Let κ_X be of order ℓ. Then $\frac{1}{\lambda}(A^{\odot\ell})^\dagger \mathrm{vec}\,(\kappa_X)$ is the vector of mixing weights $(w_1, \ldots, w_m)^T$ of the Gaussian mixture model.*

Proof From Lemma 21, $\mathrm{cum}_\ell(S_i) = \lambda w_i$. Lemma 26 implies that $\mathrm{vec}\,(\kappa_X) = \lambda A^{\odot\ell}(w_1, \ldots, w_m)^T$. Multiplying on the left by $\frac{1}{\lambda}(A^{\odot\ell})^\dagger$ gives the result. ■