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

Joseph Anderson  
Ohio State University  
andejose@cse.ohio-state.edu

Mikhail Belkin  
Ohio State University  
mbelkin@cse.ohio-state.edu

Navin Goyal  
Microsoft Research India  
navingo@microsoft.com

Luis Rademacher  
Ohio State University  
lrademac@cse.ohio-state.edu

James Voss  
Ohio State University  
vossj@cse.ohio-state.edu

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.

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 [14] 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., [15, 4, 28, 1, 16]).

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 [7] and [20], 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 [20] 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 [15] by reducing high-dimensional inference to a small number of low-dimensional problems
through appropriate projections.

However, a surprising result was recently proved in [18]. 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 [18] 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 [20] 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-
dergeneracy 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 \(O^*(e^{-\sqrt{n}})\). 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” $\sigma$. We show that if the dimension is sufficiently high, 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 27 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 [17] to obtain the necessary bounds.

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 Independent Component Analysis 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 $w_1\mathcal{N}(\mu_1, \Sigma) + \ldots + w_m\mathcal{N}(\mu_m, \Sigma)$ 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 $d \in 2\mathbb{N}$ so that $\sigma_m(B^{\odot d/2}) > 0$, then Algorithm 1 recovers each $\mu_i$ to within $\epsilon$ accuracy with probability $1 - \delta$. Its sample and time complexity are at most

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

where $w \geq \max_i(w_i)/\min_i(w_i)$, $u \geq \max_i\|\mu_i\|$, $r \geq (\max_i\|\mu_i\| + 1)/(\min_i\|\mu_i\|)$, $0 < b \leq \sigma_m(B^{\odot d/2})$ are bounds provided to the algorithm, and $\sigma = \sqrt{\lambda_{\max}(\Sigma)}$.

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 1.

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

**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 $C$,

$$\Pr\left(\sigma_{\text{min}}((M + N)^{\odot 2}) \leq \sigma^2/n^2\right) \leq 2C/n.$$
We point out the simultaneous and independent work of \cite{8}, where the authors prove learnability results related to our Theorems \ref{1} and \ref{2}. We now provide a comparison. The results in \cite{8}, 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 $\epsilon$ their algorithm has running time and sample complexity $\text{poly}_\ell(n, 1/\epsilon, 1/\rho)$ and succeeds with probability at least $1 - \exp(-Cn^{1/3})$, 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 $\text{poly}_\ell(n, 1/\epsilon, 1/\rho)$ above has degree exponential in $\ell$, unlike the degree of our bound which is polynomial in $\ell$. 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. The technique of Poissonization is, to the best of our knowledge, new in the GMM setting. Moreover, our analysis can be used in the reverse direction to obtain hardness results for ICA.

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

\textbf{Theorem 3.} Let $X$ be a set of $k^2$ 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 $C > 0$,

$$\|p - q\|_{L_1(\mathbb{R}^n)} < e^{-C(k/\log k)^{1/n}}.$$ 

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 \cite{18} for pointers). The only previous information-theoretic lower bound for learning GMMs we are aware of is due to \cite{20} and holds for two specially designed one-dimensional mixtures.

\textbf{Connection with ICA.} A key observation of \cite{18} 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 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 \textit{underdetermined ICA} setting. Two well-known algorithms for underdetermined ICA are given in \cite{10} and \cite{2}. Finally, \cite{17} provides an algorithm with rigorous polynomial time and sampling bounds for underdetermined ICA in high dimension in the presence of Gaussian noise.

\footnote{See \cite{12} Chapter 9 for a recent account of algorithms for underdetermined ICA.}
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 \( k^2 \) random \( n \)-dimensional unit vectors. Then with high probability, there exist 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 \( BS' + \eta' \) which are exponentially close as distributions in \( L^1 \) distance and satisfying: (1) The coordinate random variables of \( S \) and \( S' \) are scaled Poisson random variables. For at least one coordinate random variable, \( S_i = \alpha X_i \), where \( X \sim \text{Poisson}(\lambda) \) is such that \( \alpha \) and \( \lambda \) are polynomially bounded away from 0. (2) The Gaussian noises \( \eta \) and \( \eta' \) have polynomially bounded directional covariances.

We sketch the proof of Theorem 4 in Appendix G.

**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 polynomials in the moments of \( X \). For \( j \geq 1 \), the \( j \)th cumulant is denoted \( \text{cum}_j(X) \). Denoting \( m_j := \mathbb{E}(X^j) \), we have, for example:

\[
\text{cum}_1(X) = m_1, \text{cum}_2(X) = m_2 - m_1^2, \text{cum}_3(X) = m_3 - 3m_2m_1 + 2m_1^3.
\]

In general, cumulants can be defined as certain coefficients of a power series expansion of the logarithm of the characteristic function of \( X \):

\[
\log(\mathbb{E}(e^{itX})) = \sum_{j=1}^{\infty} \text{cum}_j(X) \frac{(it)^j}{j!}.
\]

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 \( \text{cum}_j(X + Y) = \text{cum}_j(X) + \text{cum}_j(Y) \) (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 \( \text{cum}_j(\alpha X) = \alpha^j \text{cum}_j(X) \). The first two cumulants of the standard Gaussian distribution are the mean, 0, and the variance, 1, and all subsequent Gaussian cumulants have value 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 \( e_h \) the random variable which takes on the \( i \)th canonical vector \( e_i \) with probability \( w_i \), we can write the GMM model as follows:

\[
Z = [\mu_1|\mu_2|\cdots|\mu_m]e_h + \eta.
\]

(1)

In this formulation, \( e_h \) 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 \( P S \) gives a new random vector with independent reordered coordinates, \( A P^T \) gives a new mixing matrix with reordered columns, and \( X = A P^T P S \) provides the same samples as \( X = AS \) since \( P^T P = I \).

As \( A P^T \) 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 \( \eta \) 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 \( \text{vec} (\cdot) \) from a tensor to a vector in the natural way. Namely, when and \( T \in \mathbb{R}^{n_i \times \cdots \times n_\ell} \) is a tensor, then \( \text{vec} (T)_{\delta(i_1, \ldots, i_\ell)} = T_{i_1, \ldots, i_\ell} \) where \( \delta(i_1, \ldots, i_\ell) = 1 + \sum_{j=1}^{\ell-1} n_j^{\ell-j}(i_j - 1) \) is a bijection with indices \( i_j \) running from 1 to \( n_j \). 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 [17] 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 multivariate cumulants. This tensor structure motivates the following form of the Khatri-Rao product:

**Definition 1.** Given matrices \( A \in \mathbb{R}^{a_1 \times m}, B \in \mathbb{R}^{a_2 \times m} \), a column-wise Khatri-Rao product is defined by \( A \odot B := [\text{vec} (A_1 \otimes B_1) | \cdots | \text{vec} (A_m \otimes B_m)] \), 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 \( \text{vec} (A_1 \otimes B_1) \) is flattening of the tensor \( A_1 \otimes B_1 \) into a vector. The related Khatri-Rao power is defined by \( A^{\odot \ell} = A \odot \cdots \odot A (\ell \text{ times}) \).

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 22 (Appendix H.1 from [17]) allows us to recover $A$ up to the necessary ambiguities in the noisy ICA setting. The theorem establishes guarantees for an algorithm from [17] 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 $\epsilon$, confidence parameter $\delta$, bounds on moments and cumulants $M$ and $\Delta$, a bound on the conditioning parameter $\sigma_m$, 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 $\mu_i$ up to their norms and signs, i.e. we will get $\pm \mu_i / \| \mu_i \|$.

We will then present the reduction in full. It combines the basic reduction with some preprocessing of the data to recover the $\mu_i$’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 5. Fix a positive integer $k$, and let $p_i \geq 0$ be such that $p_1 + \cdots + p_k = 1$. If $X \sim \text{Poisson}(\lambda)$ and $(Y_1, \ldots, Y_k) |_{X=x} \sim \text{Multinom}(x; p_1, \ldots, p_k)$ then $Y_i \sim \text{Poisson}(p_i \lambda)$ for all $i$ and $Y_1, \ldots, Y_k$ are mutually independent.

Basic Reduction: The main idea. Recall the GMM from equation (1) is given by $Z = [\mu_1 | \cdots | \mu_m] e_h + \eta$. Henceforth, we will set $A = [\mu_1 | \cdots | \mu_m]$. We can write the GMM in the form $Z = A e_h + \eta$, which is similar in form to the noisy ICA model, except that $e_h$ 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 $\lambda$ and $\tau$ that we will set later. We generate a Poisson random variable $R \sim \text{Poisson}(\lambda)$, 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 5 each $S_i$ has distribution $\text{Poisson}(w_i \lambda)$, 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 $\eta_i$ are iid according to $\eta_i \sim \mathcal{N}(0, \Sigma)$. In this definition, $t$ can be a random variable, in which case the $\eta_i$ 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 $\tau$, 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: (1) $X \sim AS + \eta(\tau)$, a noisy ICA model with no restriction on $R = \sum_{i=1}^{m} S_i$, and (2) $X \sim (AS + \eta(\tau))|_{R \leq \tau}$ the restricted model.

We are unable to produce samples from the first model, but it meets the assumptions of the noisy ICA problem. Pretending we have samples from model (1), we can apply Theorem 22 (Appendix H.1) to recover the Gaussian means up to sign and scaling. On the other hand, we can produce samples from model (2), and depending on the choice of $\tau$, it will be demonstrated that given an appropriate choice of $\tau$, running UnderdeterminedICA on samples from model (2) is equivalent to running UnderdeterminedICA on samples from model (1) 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 $\mu_i$ without sign or scaling ambiguities, we add an extra coordinate to the GMM as follows. The new means $\mu'_i$ are $\mu_i$ with an additional coordinate whose value is 1 for all $i$, i.e. $\mu'_i := (\mu_i^T, 1)^T$. Moreover, this coordinate has no noise. In other words, each Gaussian component now has an $(n + 1) \times (n + 1)$ covariance matrix $\Sigma' := \left( \begin{array}{cc} \Sigma & 0 \\ 0 & 0 \end{array} \right)$. 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 $u'_i := (u_i^T, 1)^T$. The reduction proceeds similarly to the above on the new inputs.

Unlike before, we will define the ICA mixing matrix to be $A' := [\mu'_1/\|\mu'_1\| \cdots |\mu'_m/\|\mu'_m\|]$ 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:

$$X' = A'S' + \eta'(\tau),$$

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

where we define $\eta'(\tau) = (\eta(\tau)^T, 0)^T$. As before, we have an ideal noisy ICA model (2) from which we cannot sample, and an approximate noisy ICA model (3) which can be made arbitrarily close to (2) in statistical distance by choosing $\tau$ appropriately. With appropriate application of Theorem 22 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 $\mu_i$: 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 $\mu_i = \frac{\tilde{A}'_i(1:n)}{\tilde{A}'_i(n+1)} \approx \frac{\tilde{A}'_i(1:n)}{A'_i(n+1)}$, with the sign indeterminacy canceling in the division.
**Subroutine 1** Single sample reduction from GMM to approximate ICA

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

**Output:** $Y$ (a sample from model (3)).

1. Generate $R$ according to $\text{Poisson}(\lambda)$.
2. If $R > \tau$ return failure.
3. Let $Y = 0$.
4. for $j = 1$ to $R$ do
   5. Get a sample $Z_j$ from the GMM.
   6. Let $Z'_j = (Z_j^T, 1)^T$ to embed the sample in $\mathbb{R}^{n+1}$.
   7. $Y = Y + Z'_j$.
8. end for
9. Let $\Sigma' = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$ (add a row and column of all zeros)
10. Generate $\eta'$ according to $\mathcal{N}(0, (\tau - R)\Sigma')$.
11. $Y = Y + \eta'$.
12. return $Y$.

### 4 Correctness of the Algorithm and Reduction

Subroutine $1$ captures the sampling process of the reduction: Let $\Sigma$ be the covariance matrix of the GMM, $\lambda$ be an integer chosen as input, and a threshold value $\tau$ also computed elsewhere and provided as input. Let $R \sim \text{Poisson}(\lambda)$. If $R$ is larger than $\tau$, 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, $\tau$ 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 (3).

Algorithm $2$ works as follows: it takes as input the parameters of the GMM (covariance matrix, number of means), tensor order (as required by $\text{UnderdeterminedICA}$), 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 $\lambda$, threshold parameter $\tau$, error parameters to be split between the “Poissonization” process and the call to $\text{UnderdeterminedICA}$, and values explicitly needed by [17] for the analysis of $\text{UnderdeterminedICA}$. 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 $\tau$ that will involve a polynomial, called $q(\Theta)$, of all the other parameters. The algorithm then calls $\text{UnderdeterminedICA}$, but instead of giving samples from the GMM, it allows access to Subroutine $1$. It is then up to $\text{UnderdeterminedICA}$ 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 $\text{UnderdeterminedICA}$ 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 the GMM.

The bounds are used instead of actual values to allow flexibility — in the context under which
Algorithm 2 Use ICA to learn the means of a GMM

**Input:** Covariance matrix $\Sigma$, 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 $\Sigma$, confidence parameter $\delta$, accuracy parameter $\epsilon$, 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).

1: Let $\delta_2 = \delta_1 = \delta/2$.
2: Let $\sigma = \sup_{v \in S_{n-1}} \sqrt{\text{Var}(v^T \eta(1))}$, for $\eta(1) \sim \mathcal{N}(0, \Sigma)$.
3: Let $\lambda = m$ be the parameter to be used to generate the Poisson random variable in Subroutine 1.
4: Let $\tau = 4 \left( \log(1/\delta_2) + \log(q(\Theta)) \right) \max \left( (e\lambda)^2, 4Cd^2 \right)$ (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 (17) in the proof of Theorem 1).
5: Let $\epsilon^* = \epsilon \left( \sqrt{1 + u^2} + 2(1 + u^2) \right)^{-1}$.
6: Let $M = \max \left( (\tau\sigma)^{d+1}, (w/\sqrt{1 + u^2})^{d+1} \right)\sigma (d + 1)^{d+1}$.
7: Let $k = d + 1$.
8: Let $\Delta = w$.
9: Invoke UnderdeterminedICA with access to Subroutine 1 parameters $\delta_1, \epsilon^*, \Delta, M$, and $k$ to obtain $\hat{A}'$ (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.
10: Divide each column of $\hat{A}'$ by the value of its last entry.
11: Remove the last row of $\hat{A}'$ to obtain $\hat{B}$.
12: return the columns of $\hat{B}$ as $\{ \tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_m \}$.


developmental comment

---

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 (2).

In the second part, we note that we do not have access to the ideal model (2), and that we can only sample from the approximate noisy ICA model (3) using the full reduction. Choosing $\tau$ 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 \( M \in \mathbb{R}^{n \times (\frac{n}{2})} \) and add a perturbation matrix \( N \in \mathbb{R}^{n \times (\frac{n}{2})} \) 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 [17], 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 \( A_k \circ \in \mathbb{R}^{(\frac{n}{2})} \), a subvector of \( A_k \circ \in \mathbb{R}^{n^2} \), given by \( (A_k \circ)_{ij} := (A_k)_i(A_k)_j \) for \( 1 \leq i < j \leq n \). Then for a matrix \( A = [A_1, \ldots, A_m] \) we have \( A \circ := [A_1 \circ, \ldots, A_m \circ] \).

**Theorem 6.** With the above notation, for any base matrix \( M \) with dimensions as above, we have, for some absolute constant \( C \),

\[
\Pr \left( \sigma_{\text{min}}((M + N) \circ) \leq \frac{\sigma^2}{n^2} \right) \leq \frac{2C}{n}.
\]

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

With probability 1, the columns of the matrix \( (M + N) \circ \) are linearly independent. This can be proved along the lines of a similar result in [17]. Fix \( k \in \binom{n}{2} \) and let \( u \in \mathbb{R}^{(\frac{n}{2})} \) be a unit vector orthogonal to the subspace spanned by the columns of \( (M + N) \circ \) 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

\[
\begin{align*}
C_k^T u &= u^T (M_k + N_k) \circ = \sum_{1 \leq i < j \leq n} u_{ij}(M_{ik} + N_{ik})(M_{jk} + N_{jk}) \\
&= \sum_{1 \leq i < j \leq n} u_{ij}M_{ik}M_{jk} + \sum_{1 \leq i < j \leq n} u_{ij}M_{ik}N_{jk} + \sum_{1 \leq i < j \leq n} u_{ij}N_{ik}M_{jk} + \sum_{1 \leq i < j \leq n} u_{ij}N_{ik}N_{jk} \\
&=: P(N_{ik}, \ldots, N_{nk}).
\end{align*}
\]

Now note that this is a quadratic polynomial in the random variables \( N_{ik} \). We will apply the anticoncentration inequality of Carbery–Wright [9] to this polynomial to conclude that the distance between the \( k' \)th column of \( (M + N) \circ \) 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 (4) becomes

\[
\text{Var}(P(N_{ik}, \ldots, N_{nk})) = \sigma^2 \left( \sum_j \left( \sum_{i < j} u_{ij}M_{ik} \right)^2 + \sum_i \left( \sum_{j < i} u_{ij}M_{ik} \right)^2 \right) + \sigma^4 \sum_{i < j} u_{ij}^2 \geq \sigma^4.
\]

In our application, our random variables \( N_{ik} \) for \( i \in [n] \) are not standard Gaussians but are iid Gaussian with variance \( \sigma^2 \), and our polynomial does not have unit variance. After adjusting for these differences using the estimate on the variance of \( P \) above, Lemma [24] gives \( \Pr(|P(N_{ik}, \ldots, N_{nk}) - \mu| \leq \epsilon) \leq 2C \sqrt{\epsilon / \sigma^2} = 2C \sqrt{\epsilon / \sigma} \).

Therefore, by the union bound over the choice of \( k \) \( \Pr(\text{dist}(C_k, C^c) \leq \epsilon) \leq (n^2)2C \sqrt{\epsilon / \sigma} \).

Now choosing \( \epsilon = \sigma^2 / n^6 \), Lemma [23] gives \( \Pr \left( \sigma_{\text{min}}((M + N) \circ) \leq \frac{\sigma^2}{n^2} \right) \leq 2C / n \). \( \square \)
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 [20] 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 \( O\left(\frac{\log n}{h}\right) \).

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: \( \mathcal{K}g(y) = \int_{\mathbb{R}^d} K(x,y)g(x)dx \). Let \( X \) be any subset of \( k \) points in \([0,1]^n\). Let \( \mathcal{K}_X \) be the kernel matrix corresponding to \( X \), \( (\mathcal{K}_X)_{ij} = K(x_i,x_j) \). It is known to be positive definite. For a function \( f : [0,1]^n \to \mathbb{R} \), the interpolant is defined as \( \mathcal{K}_X f(x) = \sum w_i K(x_i,x) \), where the coefficients \( w_i \) are chosen so that \( (\forall i) \mathcal{K}_X f(x_i) = f(x_i) \). It is easy to see that such interpolant exists and is unique, obtained by solving a linear system involving \( \mathcal{K}_X \).

We will need some properties of the Reproducing Kernel Hilbert Space \( \mathcal{H} \) corresponding to the kernel \( K \) (see [29] Chapter 10 for an introduction). In particular, we need the bound \( \|f\|_{\mathcal{H}} \leq \|f\|_{L^\infty(\mathbb{R}^n)} \) and the reproducing property, \( (f(\cdot), K(\cdot, \cdot))_H = f(x), \forall f \in \mathcal{H} \). For a function of the form \( \sum w_i K(x_i,x) \) we have \( \|\sum w_i K(x_i,x)\|_H^2 = \sum w_iw_j K(x_i,x_j) \).

**Lemma 7.** 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

\[
\|f - \mathcal{K}_X f\|_{L^\infty(\mathbb{R}^n)} < \exp\left(\frac{\log h}{h}\right).
\]

**Proof.** From [24], Theorem 6.1 (taking \( \lambda = 0 \)) we have that for some \( A > 0 \) and \( h \) sufficiently small \( \|f - \mathcal{K}_X f\|_{L^2([0,1]^n)} < \exp\left(\frac{\log h}{h}\right) \). 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 \( f - \mathcal{K}_X f \). This ultimately gives control of the norm over \( \mathbb{R}^n \) because there is a canonical isometric embedding of elements of \( \mathcal{H} \) interpreted as functions over \([0,1] \) into elements of \( \mathcal{H} \) interpreted as functions over \( \mathbb{R}^n \). We first observe that for any \( x_i \in X \), \( f(x_i) - \mathcal{K}_X f(x_i) = 0 \). Thus, from the reproducing property of RKHS, \( (f - \mathcal{K}_X f, f - \mathcal{K}_X f)_H = 0 \). Using properties of RKHS with respect to the operator \( \mathcal{K} \) (see, e.g., Proposition 10.28 of [29])

\[
\|f - \mathcal{K}_X f\|_H^2 = (f - \mathcal{K}_X f, f - \mathcal{K}_X f)_H = (f - \mathcal{K}_X f, f)_H = (f - \mathcal{K}_X f, \mathcal{K}g)_H = (f - \mathcal{K}_X f, g)_H \leq \|f - \mathcal{K}_X f\|_{L^2([0,1]^n)}\|g\|_{L^2([0,1]^n)} < \exp\left(\frac{\log h}{h}\right).
\]

Thus \( \|f - \mathcal{K}_X f\|_{L^\infty(\mathbb{R}^n)} \leq \|f - \mathcal{K}_X f\|_H < \exp\left(\frac{\log h}{h}\right). \)

\[\square\]
Theorem 8. 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 $B > 0$

$$\|p - q\|_{L^1(\mathbb{R}^n)} < \exp(B \frac{\log h}{h})$$.

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, $f = Kg$) positive function $f$ with $\int_{[0,1]} f(x)dx = 1$. Using Lemma 7 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

$$f_{X,k} - f_{Y,k} = p_1 - p_2,$$ (5)

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 (5) 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 $\frac{2}{\sqrt{\pi}} \exp(A\frac{\log h}{h})$ 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

$$\left\|\frac{1}{\alpha}p_1 - \frac{1}{\beta}p_2\right\|_{L^1(\mathbb{R})} \leq \left\|\frac{\beta}{1 - \frac{\beta}{\alpha}} p_1 - p_2\right\|_{L^1(\mathbb{R})} \leq \left|1 - \frac{\beta}{\alpha}\right| \|p_1\|_{L^1(\mathbb{R})} + \|p_1 - p_2\|_{L^1(\mathbb{R})}.$$ 

Collecting exponential inequalities completes the proof. 

af Theorem 8. 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, we see that the fill $h$ of $nm^n \log m$ points is at most $O(\sqrt{n}/m)$ with probability $1 - o(1)$. Hence, given $k$ points, we have $h = O((\log k)^{1/n})$. We see, that with a smaller probability (but still close to 1 for large $k$), we can sample $k$ points $3k^2$ times and still have the same fill.

Partitioning the set of $4k^2$ points into $2k$ disjoint subsets of $2k$ points and applying Theorem 8 (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 pigeonhole 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 $2k - 1$) 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.
References

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

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

[3] N. Alon and J. H. Spencer. *The probabilistic method*. Wiley, 2004.

[4] S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders. In *NIPS*, pages 2384–2392, 2012.

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

[6] M. Belkin, L. Rademacher, and J. Voss. Blind signal separation in the presence of Gaussian noise. In *JMLR W&CP*, volume 30: COLT, pages 270–287, 2013.

[7] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112. IEEE Computer Society, 2010.

[8] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. *CoRR*, abs/1311.3651v4, 2014.

[9] A. Carbery and J. Wright. Distributional and $L^q$ norm inequalities for polynomials over convex bodies in $R^n$. *Mathematical Research Letters*, 8:233–248, 2001.

[10] 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.

[11] 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.

[12] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation*. Academic Press, 2010.

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

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

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

[16] 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.

[17] N. Goyal, S. Vempala, and Y. Xiao. Fourier PCA. *CoRR*, http://arxiv.org/abs/1306.5825, 2013.
A. Hsu and S. M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In ITCS, pages 11–20, 2013.

M. Kendall, A. Stuart, and J. K. 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.

E. Mossel, R. O’Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: Invariance and optimality. Annals of Math., 171:295–341, 2010.

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

B. Rennie and A. Dobson. On Stirling numbers of the second kind. Journal of Combinatorial Theory, 7(2):116 – 121, 1969.

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

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

H. L. Royden, P. Fitzpatrick, and P. Hall. Real analysis, volume 4. Prentice Hall New York, 1988.

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

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

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

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

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:

$$X' = A'S' + \eta' (\tau) . \tag{6}$$

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:

$$X' = (A'S' + \eta'(\tau))|_{R \leq \tau} . \tag{7}$$
Both models rely on the choice of two parameters, $\lambda$ and $\tau$. The dependence on $\tau$ is explicit in the models. The dependence on $\lambda$ can be summarized in the unrestricted model as $S_i = \frac{1}{\|\mu_i\|} S'_i \sim \text{Poisson}(w_i, \lambda)$ independently of each other, and $R = \sum_{i=1}^{m} S_i \sim \text{Poisson}(\lambda)$.

The probability of choosing $R > \tau$ will be seen to be exponentially small in $\tau$. For this reason, running \textbf{UnderdeterminedICA} with polynomially many samples from model \textbf{(3)} will with high probability be equivalent to running the ICA Algorithm with samples from model \textbf{(7)}. 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 \textbf{(3)}. Thus, to recover the columns of $\text{ICA specialized to the ideal case}$, Theorem 11 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 $w_{\text{min}} := \min_i w_i$ and $w_{\text{max}} := \max_i w_i$. Then, the cumulants of $S'$ are bounded by the following lemma:

\textbf{Lemma 9.} Given $\ell \in \mathbb{Z}^+$, $\text{cum}_\ell(S'_i) \geq \lambda w_i \|\mu'_i\|$ for each $S'_i$. In particular, then $\text{cum}_\ell(S'_i) \geq w_{\text{min}} \lambda$.

\textbf{Proof.} By construction, $S'_i = \|\mu'_i\| S_i$. By the homogeneity property of univariate cumulants,

$$\text{cum}_\ell(S'_i) = \text{cum}_\ell(\|\mu'_i\| S_i) = \|\mu'_i\|\ell \text{cum}_\ell(S_i)$$

As $\mu'_i(n + 1) = 1$, $\|\mu'_i\| \geq 1$. The cumulants of the Poisson distribution are given in Lemma 20. It follows that $\text{cum}_\ell(S'_i) \geq \text{cum}_\ell(S_i) = \lambda w_i$. \hfill \Box

The absolute moments of Gaussian random variables are well known. For completeness, the bounds are provided in Lemma 21 of Appendix 1.

Defining $\sigma = \sup_{v \in S^{m-1}} \sqrt{\text{Var}(v^T \eta'(1))}$; vectors $\mu'_m = \max_i \|\mu'_i\|$, $\mu'_m = \min_i \|\mu'_i\|$, and similarly $\mu_m$ and $\mu_m$ for later; and choosing $\lambda = m$, we can now show a polynomial bound for the error in recovering the columns of $A'$ using \textbf{UnderdeterminedICA}.

\textbf{Theorem 11 (ICA specialized to the ideal case).} Suppose that samples of $X'$ are taken from the unrestricted ICA model \textbf{(3)} choosing parameter $\lambda = m$ and $\tau$ a constant. Suppose that \textbf{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:

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

(8)
the columns of $A'$ are recovered within error $\epsilon$ up to their signs. That is, denoting the columns returned from UnderdeterminedICA by $A'_1, \ldots, A'_m$, there exists $\alpha_1, \ldots, \alpha_m, \in \{-1, +1\}$ and a permutation $p$ of $[m]$ such that $\|A'_i - \alpha_i \hat{A}'_{\hat{p}(i)}\| < \epsilon$ 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 22. 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 22.

Parameter $d$ is fixed. We must choose $k_1, \ldots, k_m$ and $k$ such that $d < k_1 \leq k$ and cum$_{k_i}(S'_i)$ is bounded away from 0. It suffices to choose $k_1 = \cdots = k_m = k = d + 1$. By Lemma 3, 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 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 22, we have a constant

$$\Delta = \frac{w_{\min}}{w_{\max}}$$

such that 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 10, we have that

$$\mathbb{E} \left( |S'_i|^{d+1} \right) = \mathbb{E} \left( (S'_i)^{d+1} \right) \leq (\|\mu'_i\| \lambda)^{d+1} (d + 1)^{d+1}.$$  

Thus, for $M$ to bound the $(d + 1)^{th}$ moment of $S'_i$, it suffices that $M \geq (\|\mu'_{\max}\| \lambda)^{d+1} (d + 1)^{d+1}$. 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 21 that

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

$\sigma$ gives a clear upper bound for 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

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

\hspace{0.5cm} (10)

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

We can now apply Theorem 22 using the parameter values $k = d + 1$, $\Delta$ from (9), and $M$ from (10). Then with probability $1 - \delta$,

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

\hspace{0.5cm} (11)
samples suffice to recover up to sign the columns of $A'$ within $\epsilon$ accuracy. More precisely, letting $A'_1, \ldots, A'_m$ 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 $\|A'_i - A'_{p(i)}\| < \epsilon$ for each $i$.

The poly bound in (11) is equivalent to the poly bound in (8). \(\square\)

Theorem 11 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 estimates the Gaussian means from the estimates of the columns of $A'$. By propagating the errors from Theorem 11 we arrive at the following result:

**Theorem 12** (Recovery of Gaussian means in Ideal Case). Suppose that UnderdeterminedICA is run using samples of $X'$ from the ideal noisy ICA model (10) choosing parameters $\lambda = m$ and $\tau$ 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 $\tilde{A}'_1, \ldots, \tilde{A}'_m$ be the returned estimates of the columns of $A'$ (from model (10)) 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

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

samples are used, then with probability $1 - \delta$ there exists a permutation $p$ of $[m]$ such that $\|\tilde{\mu}_{p(i)} - \mu_i\| < \epsilon$ 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 11 using

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

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 $\epsilon^*$-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 $\|\alpha_i \tilde{A}'_i - \alpha A'_i\| < \epsilon^*$ holds.

This proof proceeds in two steps. First, we replace the dependencies in (13) 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 $\epsilon^*$-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 (13) when calling UnderdeterminedICA, then $A'$ will be recovered with the desired precision with probability $1 - \delta$. 

18
Claim. The $\text{poly}(\|\mu_{\max}^\prime\|^d, d^2)$ dependence in (13) can be replaced by a $\text{poly}(\|\mu_{\max}\|^d, d^2)$ dependence.

Proof of Claim. By construction, $\mu_{\max}' = \left( \begin{array}{c} \mu_{\max} \\ 1 \end{array} \right)$. By the triangle inequality,

$$\|\mu_{\max}'\|^d \leq (\|\mu_{\max}\| + 1)^d$$

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

Claim. The $\text{poly}(1/\sigma_m(A^{\odot d/2})^d)$ in (13) can be replaced by a $\text{poly}(\frac{\|\mu_{\max}\|+1}{\|\mu_{\min}\|})^d, 1/\sigma_m(B^{\odot d/2})^d$ dependence.

Proof of Claim. First define $A'$ to be the unnormalized version of $A'$. That is, $A'_i := \mu_i$. Then, $A' = A'\text{diag}(\|\mu_1\|, \ldots, \|\mu_m\|) \implies A'^{\odot d/2} = A'^{\odot d/2}\text{diag}(\|\mu_1\|^{d/2}, \ldots, \|\mu_m\|^{d/2})$. Thus, $\sigma_m(A'^{\odot d/2}) \leq \sigma_m(A^{\odot d/2}) \|\mu_{\max}'\|^{d/2}$.

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

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

Finally, we note that $B = A\text{diag}\left( \frac{1}{\|\mu_1\|}, \ldots, \frac{1}{\|\mu_m\|} \right)$ and $B^{\odot d/2} = A^{\odot d/2}\text{diag}\left( \frac{1}{\|\mu_1\|^{d/2}}, \ldots, \frac{1}{\|\mu_m\|^{d/2}} \right)$. 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:

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

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 $N > \text{poly}(1/\sigma_m(A^{\odot d/2})^d)$ samples, it suffices to have the replacement dependency of $N > \text{poly}(\frac{\|\mu_{\max}\|+1}{\|\mu_{\min}\|})^d (1/\sigma_m(B^{\odot d/2})^d) = \text{poly}(\frac{\|\mu_{\max}\|+1}{\|\mu_{\min}\|})^d 1/\sigma_m(B^{\odot d/2})^d)$ samples.

Thus, it is sufficient to call UnderdeterminedICA with

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

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

19
Step 2: Error propagation. What remains to be shown is that an appropriate choice of $\epsilon^*$ enforces $\|\mu_i - \tilde{\mu}_i\| < \epsilon$ by propagating the error.

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

$$A'_i(n+1) \geq \frac{1}{\sqrt{1 + \|\mu_i\|^2}}. \quad (15)$$

We have that:

$$\|\mu_i - \tilde{\mu}_i\| = \left\| \frac{A'_i(1:n) - \tilde{A}'_i(n+1)}{A'_i(n+1)} \right\| \leq \frac{\|A'_i(1:n) - \tilde{A}'_i(1:n)\|}{\|A'_i(n+1)\|} + \frac{\|\tilde{A}'_i(n+1)\|}{\|A'_i(n+1)\|} \frac{\|\alpha_i \tilde{A}'_i(n+1) - A'_i(n+1)\|}{\|A'_i(n+1)\|}$$

which follows in part by applying (15) for the left summand and noting that $\tilde{A}'_i$ is a unit vector for the right summand, giving the bound $\|\tilde{A}'_i(1:n)\| \leq 1$. Continuing with the restriction that $\epsilon^* < \frac{1}{2} \sqrt{1 + \|\mu_{\text{max}}\|^2}$,

$$\|\mu_i - \tilde{\mu}_i\| \leq \epsilon^* \sqrt{1 + \|\mu_{\text{max}}\|^2} + \epsilon^* \sqrt{1 + \|\mu_{\text{max}}\|^2} - \epsilon^*$$

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

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

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

which occurs when

$$\epsilon^* \leq \frac{\epsilon}{\sqrt{1 + \|\mu_{\text{max}}\|^2} + 2(1 + \|\mu_{\text{max}}\|^2)} \quad (16)$$

As $\epsilon < \frac{1}{2}$, the restriction $\epsilon^* < \frac{1}{2} \sqrt{1 + \|\mu_{\text{max}}^2\|^2}$ holds automatically for the choice of $\epsilon^*$ in (16). The sample bound from (14) contains the dependency $N > \text{poly}(\frac{1}{\epsilon^*}, \|\mu_{\text{max}}\|^2 \mu_{\text{max}}^2)$. Propagating the error gives a replacement dependency of $N > \text{poly} \left( \frac{1}{\epsilon^*}, \sqrt{1 + \|\mu_{\text{max}}\|^2}, \|\mu_{\text{max}}\|^2 \right) = \text{poly} \left( \frac{1}{\epsilon^*}, \|\mu_{\text{max}}\|^2 \right)$ as $d$ is non-negative. This propagated dependency is reflected in (12).
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 \( \tau \), this is true (c.f. Lemma 5). 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 \( \tau \) large enough that the samples still come from a distribution with arbitrarily small total variation distance to the one with truly independent coordinates.

**Lemma 13.** Fix \( \delta > 0 \). Let \( S \sim \text{Poisson}(\lambda) \) for \( \lambda \geq \ln \delta \). Let \( b = e\lambda \), If \( \tau > e\lambda \), \( \tau \geq 1 \), and \( \tau \geq \ln(1/\delta) - \lambda \), then \( \Pr(S > \tau) < \delta \).

**Proof.** By the Chernoff bound (See Theorem A.1.15 in [3]),

\[
\Pr(S > \lambda(1 + \epsilon)) \leq \left(e^\epsilon (1 + \epsilon)^{(1+\epsilon)})^\lambda.
\]

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

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

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

\[
\tau(1 - \log_b \tau) = \tau - \tau \log_b \tau = \log_b(b^\tau (1/\tau)^\tau).
\]

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

\[
\log_b(b^\tau (1/\tau)^\tau) \leq \log_b(\delta e^{\lambda})
\]

which then implies it suffices that

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

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

**Lemma 14.** Let \( N, \delta > 0 \), \( N \in \mathbb{N} \), and \( T_1, T_2, \ldots, T_N \) be iid with distribution \( \text{Poisson}(\lambda) \). If \( \tau \geq \ln(N/\delta) - \lambda \) then

\[
\Pr \left( \bigcup_i \{ T_i > \tau \} \right) < \delta.
\]

**Proof.** By Lemma 13 \( \tau \geq \ln(N/\delta) - \lambda \) implies \( \Pr(T_i > \tau) < \delta/N \) for every \( i \). The union bound gives us the desired result. \( \square \)

It should now be easy to see that if we choose our threshold \( \tau \) 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 15.** Fix \( \delta > 0 \). Let \( \tau > 0 \). Let \( F \) be a Poisson distribution with parameter \( \lambda \) 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 \( d_{TV}(F,G) = 1 - F(\tau) \).
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

\[ d_{TV}(F, G) = \frac{|F(\tau) - 1|}{2F(\tau)} \sum_{i=0}^{\tau} f(i) + \frac{1}{2} \sum_{i=\tau+1}^{\infty} f(i) = \frac{|F(\tau) - 1|}{2} + \frac{1 - F(\tau)}{2} = 1 - F(\tau). \]

\[ \square \]

A.3 Proof of Theorem 1

We now show that after the reduction is applied, we can use the UnderdeterminedICA routine given in [17] 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. The algorithm is provided parameters: Covariance matrix \( \Sigma \), upper bound on tensor order \( d \), access to samples from a mixture of \( m \) identical spherical Gaussians in \( \mathbb{R}^n \) with covariance \( \Sigma \), confidence \( \delta \), accuracy \( \epsilon \), 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 \( r \geq (\max_i \|\mu_i\| + 1)/(\min_i \|\mu_i\|) \).

The algorithm then needs to fix the number of samples \( N \), sampling threshold \( \tau \), Poisson parameter \( \lambda \), and two new errors \( \delta_1 \) and \( \delta_2 \) so that \( \delta_1 + \delta_2 \leq \delta \). For simplicity, we will take \( \delta_1 = \delta_2 = \delta/2 \). Then fix \( \sigma = \sup_{u \in S^{n-1}} \sqrt{\text{Var}(v^T \eta(1))} \) for \( \eta(1) \sim \mathcal{N}(0, \Sigma) \). Recall that \( B \) is the matrix whose \( i \)th column is \( \mu_i/\|\mu_i\| \). Let \( A' \) be the matrix whose \( i \)th 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 (2)) and the mean matrix \( B \) satisfies \( \sigma_m(B^{\odot d/2}) \geq b > 0 \). Then by Theorem 12, with probability of error \( \delta_1 \), the call to UnderdeterminedICA in Algorithm 2 recovers the columns of \( B \) to within \( \epsilon \) and up to a permutation using \( N \) samples of complexity

\[ p(\tau^{d^2}, \Theta) = \text{poly}(n^d, m^{d^2}, (\tau\sigma)^d, u^{d^2}, w^{d^2}, d^{d^2}, r^{d^2}, 1/b^d, 1/\epsilon, 1/\delta_1) \]

where \( p(\tau^{d^2}, \Theta) \) is the bound on \( N \) promised by Theorem 12 and \( \Theta \) is all its arguments except the dependence in \( \tau \). 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 \( \epsilon \) 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 \( R \sim \text{Poisson}(\lambda) \). Given \( \delta' \), Lemma 15 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 \( \text{Poisson}(\lambda) \) distribution. Let \( G \) be a distribution given by density function \( g(x) = (f(x)1_{0 \leq x \leq \tau})/F(\tau) \). 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,

\[
d_{TV}(F', G') \leq \sum_{i=1}^{N} d_{TV}(F, G) = Nd_{TV}(F, G).
\]

Then for our choice of \( \tau \), by Lemma 13 and Lemma 15 we have

\[
d_{TV}(F', G') \leq Nd_{TV}(F, G) = NPr(X_1 > \tau) \leq N \delta' = \delta_2/2.
\]

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 \( \tau \) as \( p(\tau^{d^2}, \Theta) \). This means there will be a power of \( \tau \) which dominates all of the \( \tau \) factors in \( p \), and in particular, will be \( \tau^{Cd^2} \) for some \( C \). It then suffices to choose \( C \) so that \( p(\tau^{d^2}, \Theta) \leq \tau^{Cd^2} q(\Theta) \leq N \), where

\[
q(\Theta) = \text{poly} \left( n^d, m^d, \sigma^d, u^d, w^d, d^d, r^d, 1/b^d, 1/\epsilon, 1/\delta_1 \right).
\]

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

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

Since \( \lambda \geq 1 \) it suffices to choose \( \tau \) so that

\[
\frac{2}{\delta} q(\Theta) \tau^{Cd^2} \leq \frac{\tau^\tau}{\tau^{Cd^2} (e\lambda)^\tau}.
\]

Finally, we claim that

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

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 \( x^d \leq x^e/b^d \). 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 \( \tau \), to satisfy (18), 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 \( \tau^{\tau/4} \geq \frac{\tau^{\tau/4}}{2} \log q(\Theta) \) which implies \( \tau^{\tau/4} \geq \frac{e^{\tau/4}}{2} \leq \frac{2}{\delta} q(\Theta) \). Thus for our choice of \( \tau \), 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

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

where we used the bound \( q(\Theta) \leq (n^d m^d \sigma^d u^d w^d (d + 1)^d r^d/b^d \delta_{1,\epsilon})^{O(1)} \). By the choice of \( \tau \), one can absorb \( \tau^{d^2} \) into the above \( \text{poly}(\cdot) \) expression, giving the result. \qed
B Lemmas on the Poisson Distribution

The following lemmas are well-known; see, e.g., [13]. We provide proofs for completeness.

**Lemma 16.** If $X \sim \text{Poisson}(\lambda)$ and $Y \mid X=x \sim \text{Bin}(x, p)$ then $Y \sim \text{Poisson}(p\lambda)$.

**Proof.**

\[
\Pr(Y = y) = \sum_{x:y \geq y} \Pr(Y = y \mid X = x) \Pr(X = x) \\
= \sum_{x:y \geq y} \binom{x}{y} p^y (1-p)^{x-y} \frac{\lambda^x e^{-\lambda}}{x!} \\
= p^y e^{-\lambda} \sum_{x:y \geq y} \frac{\lambda^x}{x!} \binom{x}{y} (1-p)^{x-y} \\
= \frac{(p\lambda)^y e^{-\lambda}}{y!} \sum_{x:y \geq y} \frac{(\lambda(1-p))^{x-y}}{(x-y)!} \\
= \frac{(p\lambda)^y e^{-\lambda}}{y!} e^{(1-p)\lambda} \\
= \frac{(p\lambda)^y e^{-p\lambda}}{y!}.
\]

\[\square\]

**Lemma 17.** Fix a positive integer $k$, and let $p_i \geq 0$ be such that $p_1 + \cdots + p_k = 1$. If $X \sim \text{Poisson}(\lambda)$ and $(Y_1, \ldots, Y_k) \mid X=x \sim \text{Multinom}(x; p_1, \ldots, p_k)$ then $Y_i \sim \text{Poisson}(p_i\lambda)$ for all $i$ and $Y_1, \ldots, Y_k$ are mutually independent.

**Proof.** The first part of the lemma (i.e., $Y_i \sim \text{Poisson}(p_i\lambda)$ for all $i$) follows from Lemma 16. For the second part, let’s prove it for the binomial case ($k = 2$); the general case is similar.

\[
\Pr(Y_1 = y_1, Y_2 = y_2) = \Pr(Y_1 = y_1, Y_2 = y_2 \mid X = y_1 + y_2) \Pr(X = y_1 + y_2) \\
= \binom{y_1 + y_2}{y_1} p^{y_1}(1-p)^{y_2} \cdot \frac{\lambda^{y_1+y_2} e^{-\lambda}}{(y_1 + y_2)!} \\
= \frac{(p\lambda)^{y_1} e^{-p\lambda}}{y_1!} \cdot \frac{(1-p)\lambda^{y_2} e^{-(1-p)\lambda}}{y_2!} \\
= \Pr(Y_1 = y_1) \cdot \Pr(Y_2 = y_2).
\]

\[\square\]

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 $\sigma$ give a permutation of $k$ indices. Then, $\kappa_{Y^i_1, \ldots, i^k_k} = \kappa_{Y^{\sigma(i_1), \ldots, \sigma(i_k)}}$. 

24
• (Multilinearity of coordinate random variables) Given constants $\alpha_1, \cdots, \alpha_\ell$, then
\[
\text{cum}(\alpha_1 Y_{i_1}, \cdots, \alpha_\ell Y_{i_\ell}) = \left(\prod_{i=1}^\ell \alpha_i\right) \text{cum}(Y_{i_1}, \cdots, Y_{i_\ell}).
\]
Also, given a scalar random variable $Z$, then
\[
\text{cum}(Y_{i_1} + Z, Y_{i_2}, \cdots, Y_{i_\ell}) = \text{cum}(Y_{i_1}, Y_{i_2}, \cdots, Y_{i_\ell}) + \text{cum}(Z, Y_{i_2}, \cdots, Y_{i_\ell})
\]
with symmetry implying the additive multilinear property for all other coordinates.

• (Independence) If there exist indices $i_j, i_k$ such that $Y_{i_j}$ and $Y_{i_k}$ are independent random variables, then the cross-cumulant $\kappa_{Y_{i_1}, \cdots, Y_{i_\ell}} = 0$. Combined with multilinearity, it follows that when there are two independent random vectors $Y$ and $Z$, then $\kappa_{Y + Z, Y_{i_1} + \cdots + Y_{i_\ell}} = \kappa_Y + \kappa_Z$.

• (Vanishing Gaussians) When $\ell \geq 3$, then for the Gaussian random variable $\eta$, $\kappa_\eta = 0$.

## D Bounds on Stirling Numbers of the Second Kind

The following bound comes from [23, Theorem 3].

**Lemma 18.** If $n \geq 2$ and $1 \leq r \leq n - 1$ are integers, then \[ \left\{ \begin{array}{l} n \cr r \end{array} \right\} \leq \frac{1}{2} \left( \begin{array}{l} n \cr r \end{array} \right) r^{n-r}. \]

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

**Lemma 19.** If $n, r \in \mathbb{Z}^+$ such that $r \leq n$, then \[ \left\{ \begin{array}{l} n \cr r \end{array} \right\} \leq n^{n-1}. \]

**Proof.** The Stirling number \[ \left\{ \begin{array}{l} n \cr k \end{array} \right\} \] 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 \[ \left\{ \begin{array}{l} n \cr r \end{array} \right\} = 1 \]
As $n \geq 1$, it is clear that for these choices of $n$ and $r$, \[ \left\{ \begin{array}{l} n \cr r \end{array} \right\} \leq n^{n-1}. \]
By the restriction $1 \leq r \leq n$, when $n = 1$, then $n = r$ giving that \[ \left\{ \begin{array}{l} n \cr 1 \end{array} \right\} = 1. \]
As such, the only remaining cases to consider are when $n \geq 2$ and $1 \leq r \leq n - 1$, the cases where Lemma 18 applies.

When $n \geq 2$ and $1 \leq r \leq n - 1$, then
\[
\left\{ \begin{array}{l} n \cr r \end{array} \right\} \leq \frac{1}{2} \left( \begin{array}{l} n \cr r \end{array} \right) r^{n-r} = \frac{1}{2} \frac{n!}{r!(n-r)!} r^{n-r} \leq \frac{1}{2} n^r r^{n-r-1} < \frac{1}{2} n^r n^{n-r-1} = \frac{1}{2} n^{n-1},
\]
which is slightly stronger than the desired upper bound. \qed

## 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 20** (Cumulants of the Poisson distribution). Let $X \sim \text{Poisson}(\lambda)$. Then, $\text{cum}_\ell(X) = \lambda$ for every positive integer $\ell$. 

25
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 $\ell$th derivative ($\ell \geq 1$) is given by $g^{(\ell)}(t) = \lambda e^t$.

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

Lemma 21 (Absolute moments of the Gaussian distribution). The absolute moments of the Gaussian random variable $\eta \sim N(0, \sigma^2)$ are given by:

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

The case that $\ell$ is even in Lemma 21 is well known, and can be found for instance in [19, Section 3.4]. For general $\ell$, it is known (see [30]) that

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

When $\ell$ is odd, $\frac{\ell+1}{2}$ is an integer, allowing the Gamma function to simplify to a factorial: $\Gamma\left(\frac{\ell+1}{2}\right) = \left(\frac{\ell}{2}\right)!$. This gives the case where $\ell$ is odd in Lemma 21.

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 $\Omega$ with sigma-algebra $\Sigma$, the total variation is denoted and defined as:

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

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

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

where $\mu$ 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

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

where we choose $\mu$ that simply assigns unit measure to each atom of $\Omega$ (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 [22] and Sect. 11.6 in [26].
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 \ref{thm:lower_bound} gives us two GMMs $p$ and $q$ of identity covariance Gaussians that are exponentially close with respect to $k^2$ (the number of points used to generate the Gaussian means) in $L^1$ distance but far in parameter distance. We apply the basic reduction from Section \ref{sec:reduction} with $\lambda$ 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 $X_q = A_q S_q + \eta(\tau)$ (model (1) from Section \ref{sec:reduction}). Then, we let $S_p$ and $S_q$ take on the scaling information of the ICA model by replacing $S_{pi}$ and $S_{qj}$ by $\|A_p\| S_{pi}$ and $\|A_q\| S_{qj}$ respectively, and replacing the columns of $A_p$ and $A_q$ with their unit-normalized versions. While Theorem \ref{thm:lower_bound} is proven in the setting where Gaussian means are drawn uniformly at random from the hypercube, it can be reformulated to have Gaussian means drawn uniformly at random from the unit ball. Under such a reformulation, the columns of $A_q$ and $B_q$ are chosen from a set of $k^2$ points taken uniformly from the unit sphere $S^{n-1} \subset \mathbb{R}^n$, which are thus well separated with high probability.

Recall that $R_p = \sum_i S_{pi}$ and $R_q = \sum_i S_{qj}$ are Poisson distributed with parameters $m_p$ and $m_q$ denoting the number of columns of $A_p$ and $A_q$ respectively. Lemma \ref{lem:poisson} implies that for a choice of $\tau_p$ which is linear in $m_p$, the probability of a draw with $R_p > \tau_p$ is exponentially small, and similarly for $\tau_q$. In particular, we choose $\tau = \max(\tau_p, \tau_q)$ for the above ICA models.

Now since the $L^1$ (and hence total variation) distance between $p$ and $q$ is exponentially small in $k^2$ (upper bound on the number of components), the distance between the two resulting ICA models produced by the reduction is also exponentially small (specifically, the total variation distance between the random variables $X_p$ and $X_q$). 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, and hence its contribution to the overall total variation distance between $X_p$ and $X_q$ is 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 (that is for random variables $C, D, E, F$, $d_{TV}(C + D, E + F) \leq d_{TV}(C, E) + d_{TV}(D, F)$) 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, 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 in $k$ in total variation distance. In particular, the sample complexity of distinguishing between $X_p$ and $X_q$ is exponential in $k$.

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 Guassian noise, then by defining $A' := [A|I_n]$, and $S' := [ST, \eta^T]^T$ 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 \cite{le2017}) 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).
Theorem 22 ([17]). Let a random vector $x \in \mathbb{R}^n$ be given by an underdetermined ICA model with unknown Gaussian noise $x = As + \eta$ where $A \in \mathbb{R}^{n \times m}$ 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 $\sigma_m(A^{\circ d/2}) > 0$. Let $k > d$ be such that for each $s_i$, there is a $k_i$ satisfying $d < k_i \leq k$ and $|\text{cum}_{k_i}(s_i)| \geq \Delta$, and $E\left(|s_i|^k\right) \leq M$. Moreover, suppose that the noise also satisfies the same moment condition: $E\left(|u, \eta|^k\right) \leq M$ for any unit vector $u \in \mathbb{R}^n$ (this is satisfied if we have $k!\sigma^k \leq M$ where $\sigma^2$ is the maximum eigenvalue of $\Sigma$). Then algorithm \textbf{UnderdeterminedICA} returns a set of $n$-dimensional vectors $(\tilde{A}_i)_{i=1}^m$ so that for some permutation $\pi$ 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 $\text{poly}\left(n^k, m^k, M^k, 1/\Delta^k, 1/\sigma_m(A^{\circ d/2})^k, 1/\epsilon, 1/\delta\right)$.

H.2 Rudelson-Vershynin subspace bound

Lemma 23 (Rudelson–Vershynin [27]). If $A \in \mathbb{R}^{n \times m}$ has columns $C_1, \ldots, C_m$, then denoting $C_{-i} = \text{span}(C_j : j \neq i)$, we have

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

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

H.3 Carbery-Wright anticoncentration

The version of the anticoncentration inequality we use is explicitly given in [21] which in turn follows immediately from [9]:

Lemma 24 ([21]). Let $Q(x_1, \ldots, x_n)$ be a multilinear polynomial of degree $d$. Suppose that $\text{Var}(Q) = 1$ 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 $\epsilon > 0$,

$$\Pr_{(x_1, \ldots, x_n) \sim \mathcal{N}(0, I_n)}(|Q(x_1, \ldots, x_n) - t| \leq \epsilon) \leq C\epsilon^{1/d}.$$
are defined as the coefficients of the Taylor expansion of \( g_Y(t) \) around 0, and can be obtained using the formula

\[
\kappa^{j_1 \ldots j_\ell}_Y = \frac{\partial}{\partial j_1} \cdots \frac{\partial}{\partial j_\ell} g_Y(t) \bigg|_{t=0}.
\]

When unindexed, \( \kappa_Y \) will denote the full order-\( \ell \) 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 \( \text{cum}_\ell(Y_j) = \kappa^{j \ldots j}_Y \) (\( j \) repeated \( \ell \) 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 [11]. In the fully determined ICA setting in which the number source signals does not exceed the ambient dimension, the papers [4] and [6] 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 [10] and later in [2] to solve underdetermined ICA. Finally, [17] 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 \( \lambda \), we have that \( X = AS + \eta \) is observed such that \( A = [\mu_1| \cdots |\mu_m] \) and \( S_i \sim \text{Poisson}(w_i \lambda) \). As \( A \) has already been recovered, what remains to be recovered are the weights \( w_1, \ldots, w_m \). These can be recovered using the tensor structure of higher order cumulants. The critical relationship is captured by the following Lemma:

**Lemma 25.** Suppose that \( X = AS + \eta \) gives a noisy ICA model. When \( \kappa_X \) is of order \( \ell > 2 \), then

\[
\text{vec} (\kappa_X) = A^\otimes \ell (\text{cum}_\ell(S_1), \ldots, \text{cum}_\ell(S_m))^T.
\]

**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 \( \kappa_X \):

\[
\kappa^{i_1 \ldots i_\ell}_X = \kappa^{i_1 \ldots i_\ell}_{AS} = \text{cum}((AS)_{i_1, \cdots, (AS)_{i_\ell})
\]

\[
= \text{cum} \left( \sum_{j_1=1}^m A_{i_1 j_1} S_{j_1} \cdots, \sum_{j_\ell=1}^m A_{i_\ell j_\ell} S_{j_\ell} \right)
\]

\[
= \sum_{j_1, \ldots, j_\ell \in [m]} \left( \prod_{k=1}^\ell A_{i_k j_k} \right) \text{cum}(S_{j_1}, \ldots, S_{j_\ell}) \quad \text{by multilinearity}
\]

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

\[
\kappa^{i_1 \ldots i_\ell}_X = \sum_{j=1}^m \left( \prod_{k=1}^\ell A_{i_k j_k} \right) \text{cum}_\ell(S_j) = \sum_{j=1}^m (A_j)^\otimes \ell_{i_1, \ldots, i_\ell} \text{cum}_\ell(S_j)
\]

Flattening yields:

\[
\text{vec} (\kappa_X) = A^\otimes \ell (\text{cum}_\ell(S_1), \ldots, \text{cum}_\ell(S_m))^T.
\]

In particular, we have that \( S_i \sim \text{Poisson}(w_i \lambda) \) 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 26. 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 $\kappa_X$ be of order $\ell$. Then $\frac{1}{\lambda} (A^{\odot \ell})^\dagger \text{vec} (\kappa_X)$ is the vector of mixing weights $(w_1, \ldots, w_m)^T$ of the Gaussian mixture model.

Proof. From Lemma 20, $\text{cum}_\ell(S_i) = \lambda w_i$. Lemma 25 implies that $\text{vec} (\kappa_X) = \lambda A^{\odot \ell} (w_1, \ldots, w_m)^T$. Multiplying on the left by $\frac{1}{\lambda} (A^{\odot \ell})^\dagger$ gives the result. \qed