Beyond Parallel Pancakes: Quasi-Polynomial Time Guarantees for Non-Spherical Gaussian Mixtures

Rares-Darius Buhai RARES.BUHAI@INF.ETHZ.CH

David Steurer DSTEURER@INF.ETHZ.CH

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider mixtures of $k \geq 2$ Gaussian components with unknown means and unknown covariance (identical for all components) that are well-separated, i.e., distinct components have statistical overlap at most k^{-C} for a large enough constant $C \geq 1$.

Previous statistical-query Diakonikolas et al. (2017) and lattice-based Bruna et al. (2021); Gupte et al. (2022) lower bounds give formal evidence that, even for the special case of colinear means, distinguishing such mixtures from (pure) Gaussians may be exponentially hard (in k).

We show that, surprisingly, this kind of hardness can only appear if mixing weights are allowed to be exponentially small. For polynomially lower bounded mixing weights, we show how to achieve non-trivial statistical guarantees in quasi-polynomial time.

Concretely, we develop an algorithm based on the sum-of-squares method with running time quasi-polynomial in the minimum mixing weight. The algorithm can reliably distinguish between a mixture of $k \geq 2$ well-separated Gaussian components and a (pure) Gaussian distribution. As a certificate, the algorithm computes a bipartition of the input sample that separates some pairs of mixture components, i.e., both sides of the bipartition contain most of the sample points of at least one component.

For the special case of colinear means, our algorithm outputs a k-clustering of the input sample that is approximately consistent with all components of the underlying mixture. We obtain similar clustering guarantees also for the case that the overlap between any two mixture components is lower bounded quasi-polynomially in k (in addition to being upper bounded polynomially in k).

A significant challenge for our results is that they appear to be inherently sensitive to small fractions of adversarial outliers unlike most previous algorithmic results for Gaussian mixtures. The reason is that such outliers can simulate exponentially small mixing weights even for mixtures with polynomially lower bounded mixing weights.

A key technical ingredient of our algorithms is a characterization of separating directions for well-separated Gaussian components in terms of ratios of polynomials that correspond to moments of two carefully chosen orders logarithmic in the minimum mixing weight.

1. Introduction

Gaussian mixture models (GMMs) are among the most extensively studied statistical models in a wide range of scientific disciplines Pearson (1894); Dasgupta (1999); Ashtiani et al. (2020). Over the course of the last two decades, a major body of research explored what kinds of algorithmic guarantees are feasible for GMMs Dasgupta (1999); Vempala and Wang (2002); Kalai et al. (2010); Moitra and Valiant (2010); Hsu and Kakade (2013).

Recent years have seen significant algorithmic advances along two dimensions.

The first kinds of advances concern mixtures of a large number of spherical Gaussians, i.e., Gaussians with identity I_d as covariance.[1] Several works showed how to cluster such mixtures in time quasi-polynomial in the number k of components under a minimum mean-separation requirement of $O(\sqrt{\log k})$, which up to a constant factor matches the minimum separation that guarantees clusterability of the mixture Hopkins and Li (2018); Diakonikolas et al. (2018); Kothari et al. (2018). In a recent breakthrough, the running time has been improved to polynomial assuming a slightly larger minimum separation of $O(\log^{1/2+c} k)$ for any $c > 0$ Liu and Li (2022). Even without any separation requirement, it is possible to compute quasi-polynomially sized covers of the set of means Diakonikolas and Kane (2020).

The second kinds of advances concern mixtures of a small number of Gaussian components with unknown covariances. These advances extended previous algorithmic guarantees to the robust setting, i.e., in the presence of a small constant fraction of adversarially chosen outliers. Concretely, it is now possible to estimate the parameters of an arbitrary mixture of Gaussian components in the presence of such outliers Bakshi et al. (2020a); Liu and Moitra (2021); Bakshi et al. (2020b). The running time is polynomial in the ambient dimension but (at least) exponential in the number of components.

One of the most outstanding challenges remaining in this area is to clarify what kinds of algorithmic guarantees are possible when the number of components is large and their covariances are unknown. So far, mixtures of a large number of Gaussian components with unknown covariances have defied comparable algorithmic progress.[2] Indeed, there is formal evidence, in the form of statistical-query Diakonikolas et al. (2017) and lattice-based Bruna et al. (2021); Gupte et al. (2022) lower bounds, to suggest that this setting is computationally inherently harder than the spherical setting. Specifically, these results suggest that even for k components with tiny statistical overlaps, say at most 2^{-k}, approximately clustering the components may require time exponential in k despite the sample complexity being polynomial in d and k. Underlying this evidence is the well-known parallel pancakes construction: Orthogonal to a randomly chosen direction u, all components of this mixture distribution agree with a (pure) standard Gaussian distribution, and along direction u, the components are well-separated but their mixture matches the first k moments of a (univariate) standard Gaussian distribution.[3]

In this work, we show that, surprisingly, this kind of hardness can appear only in the case that mixing weights are allowed to be exponentially small. Indeed, we develop algorithms with substantial statistical guarantees that run in quasi-polynomial time whenever the mixing weights are bounded from below by a polynomial. Before our work, the best known running times to achieve these kinds of guarantees were (at least) exponential in k. We hope that our work opens up

[1] Many known algorithms for mixtures of Gaussians with covariance I_d also extend to somewhat more general settings, e.g., the case that the covariances are different multiples of I_d or diagonal matrices (axis-aligned case) or that case that the covariance is upper bounded in the Loewner order by I_d. For our discussion, we focus on the simplest case (all covariances identity) because, to the best of our knowledge, these kinds of generalizations are orthogonal to the kind of generalization we aim for in this work.

[2] A notable exception is a particular smoothed model for such mixtures when the ambient dimension is large enough Ge et al. (2015).

[3] As a consequence of this construction, all means are colinear with u. We also emphasize that these means are well-separated relative to the variance of each component in direction u. However, since this variance is very small (about $k^{-O(1)}$), the standard Euclidean distance between the components is tiny. We remark that parallel pancakes constructions have been discussed in the literature already before Diakonikolas et al. (2017). The influential work Brubaker and Vempala (2008) provided an efficient algorithm for the case $k = 2$.

a new direction of research on efficient algorithms for mixtures of well-separated Gaussians with polynomially lower bounded mixing weights.

Within this new direction of research, we identify the following appealing open question:

> *Consider a mixture of $k \geq 2$ Gaussian components with unknown means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and unknown covariance $\Sigma \in \mathbb{R}^{d \times d}$ (identical for all components)* [4] *and with minimum mixing weight $p_{\min} > 0$. Suppose the components are well-separated in the sense that any two distinct components have statistical overlap at most $p_{\min}^C$ for a large enough constant $C \geq 1$.*
>
> *Given a sample of size $n \geq d^{O(\log(1/p_{\min}))}$, can we compute in time polynomial in n a k-clustering of the sample that is consistent with the mixture components on all but at most a $p_{\min}^{10}$ fraction of the sample?*

We conjecture that such an algorithm does exist. Indeed, we confirm the conjecture for the special case that the means are colinear (Theorem 2) and under a diameter bound (Theorem 3). In the general case, our algorithm provides a somewhat weaker guarantee and computes only a bipartition of the sample that separates at least one pair of mixture components (Theorem 1).[5]

We identify an interesting challenge in the context of establishing the above conjecture that our techniques can partially overcome: Any hypothetical algorithm establishing the above conjecture or our (non-hypothetical) algorithms inherently cannot be robust to even a tiny fraction of outliers (assuming the hardness of the parallel-pancakes constructions in Diakonikolas et al. (2017); Bruna et al. (2021); Gupte et al. (2022)). The reason is that a tiny $1/k^{100}$ fraction of outliers are enough to simulate these hard instances by adding components with appropriately decaying mixing weights and spaced means. At the same time, many recent algorithmic approaches in the context of GMMs are inherently tied to robustness. For example, certain kinds of identifiability proofs used in the analysis of sum-of-squares based algorithms automatically imply robust algorithms. Also many kinds of iteration schemes inherently require robustness for their subroutines in order to guarantee that the next iteration can successfully deal with the errors introduced by previous iterations.

1.1. Results

Separating bipartition Suppose we are given a quasi-polynomial size sample of a mixture of k Gaussian components with unknown means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and unknown covariance $\Sigma \in \mathbb{R}^{d \times d}$ and with minimum mixing weight at least $1/k^{100}$ such that there exists a pair of mixture components $a \neq b$ with $\|\Sigma^{-1/2}(\mu_a - \mu_b)\| \gg \sqrt{\log k}$. Then, as Theorem 1 shows, it is possible to compute in quasi-polynomial time a bipartition of the samples such that, for each side of the bipartition, there exists a component with 0.99 of its samples assigned to it.

Theorem 1 *Given a sample of size $n \geq (kd)^{O(\log k)}$ from a mixture of k Gaussian components $N(\mu_1, \Sigma)$,…,$N(\mu_k, \Sigma)$ with minimum mixing weight at least $1/k^{100}$ such that $\max_{a \neq b} \|\Sigma^{-1/2}(\mu_a -$*

[4] The hard instances in all lower bounds cited above are mixtures with identical covariances. A natural motivation to consider identical covariances is affine invariance. Algorithms for the spherical case assume that the input data is presented in a favorable affine transformation. However, a natural property desired for algorithms operating on geometric data is to be invariant under affine transformations Brubaker and Vempala (2008).

[5] The algorithm for the general case also computes such a bipartition under the weaker assumption that *there exists* a pair of mixture components that has small overlap (as opposed to all pairs having small overlap). Under this assumption full clustering is impossible and a partial clustering seems the appropriate guarantee to aim for.

$\mu_b)\| \gg \sqrt{\log k}$, there exists an algorithm that runs in time $n \cdot d^{O(\log k)}$ and returns with probability 0.99 a partition of $[n]$ into two sets C_1 and C_2 such that, if the true clustering of the samples is $S_1, ..., S_k$, then

$$\max_i \frac{|C_1 \cap S_i|}{|S_i|} \geq 0.99 \quad \textit{and} \quad \max_i \frac{|C_2 \cap S_i|}{|S_i|} \geq 0.99\,.$$

For general mixing weights, the same result holds with k replaced by $1/p_{\min}$ in all guarantees. See Theorem 11 for the full result.

Colinear means Suppose, in addition, that the mixture of Guassians is well-separated, i.e., the minimum mean separation satisfies $\min_{a \neq b} \|\Sigma^{-1/2}(\mu_a - \mu_b)\| \gg \sqrt{\log k}$, and that the unknown means $\mu_1, ..., \mu_k$ are colinear. Given a quasi-polynomial number of samples from the mixture, Theorem 2 shows that it is possible to compute in quasi-polynomial time a partition of the samples into k clusters such that the fraction of samples assigned to incorrect clusters is polynomially small in k.

For simplicity, in the theorem statement below we assume that all eigenvalues of Σ and all eigenvalues of the covariance matrix of the mixture are polynomially lower and upper bounded in k and d.

Theorem 2 *Given a sample of size $n \geq (kd)^{O(\log k)}$ from a mixture of k Gaussian components $N(\mu_1, \Sigma), \ldots, N(\mu_k, \Sigma)$ with minimum mixing weight at least $1/k^{100}$ such that $\min_{a \neq b} \|\Sigma^{-1/2}(\mu_a - \mu_b)\| \gg \sqrt{\log k}$ and $\mu_1, ..., \mu_k$ colinear, there exists an algorithm that runs in time $n^{O(\log k)}$ and returns with high probability a partition of $[n]$ into k sets $C_1, ..., C_k$ such that, if the true clustering of the samples is $S_1, ..., S_k$, then there exists a permutation π of $[k]$ such that*

$$1 - \frac{1}{n} \sum_{i=1}^{k} |C_i \cap S_{\pi(i)}| \leq k^{-O(1)}\,.$$

For general mixing weights, the same result holds with k replaced by $1/p_{\min}$ in all guarantees. See Theorem 39 for the full result.

Given such a clustering, we can also recover the means and the covariance of the components using robust Gaussian estimation algorithms Diakonikolas et al. (2019) or robust moment estimation algorithms Kothari et al. (2018). For example, via Kothari et al. (2018), we obtain a multiplicative approximation to the covariance $(1 - k^{-O(1)})\Sigma \preceq \hat{\Sigma} \preceq (1 + k^{-O(1)})\Sigma$ and a "covariance-aware" approximation to the means $\|\Sigma^{-1/2}(\hat{\mu}_i - \mu_i)\| \leq k^{-O(1)}$.

Small radius If instead of colinear means we have bounded means $\|\Sigma^{-1/2}\mu_i\| \leq R$ with $R = \mathrm{polylog}(k)$, Theorem 3 shows that it is again possible to cluster the samples with a quasi-polynomial number of samples and quasi-polynomial time.

Theorem 3 *Given a sample of size $n \geq (kd)^{O(R^2 + \log k)}$ from a mixture of k Gaussian components $N(\mu_1, \Sigma), \ldots, N(\mu_k, \Sigma)$ with minimum mixing weight at least $1/k^{100}$ such that $\min_{a \neq b} \|\Sigma^{-1/2}(\mu_a - \mu_b)\| \gg \sqrt{\log k}$ and $\|\Sigma^{-1/2}\mu_i\| \leq R$, there exists an algorithm that runs in time $n^{O(R^2 + \log k)}$ and returns with high probability a partition of $[n]$ into k sets $C_1, ..., C_k$ such that, if the true clustering of the samples is $S_1, ..., S_k$, then there exists a permutation π of $[k]$ such that*

$$1 - \frac{1}{n} \sum_{i=1}^{k} |C_i \cap S_{\pi(i)}| \leq k^{-O(1)}\,.$$

For general mixing weights, the same result holds with k replaced by $1/p_{\min}$ in all guarantees. See Theorem 47 for the full result.

As in the case of colinear means, given such a a clustering, we can recover the means and the covariance of the components.

Unlike our results for separating bipartitions and colinear means, this result follows from a direct reduction to a previous algorithm for spherical components Hopkins and Li (2018). Concretely, we observe that this algorithm requires only a rough multiplicative approximation (in the SOS sense) of a polynomial of the form $q(v) = \|\Sigma^{1/2}v\|^t$ for some t polylogarithmic in k. As we show, the empirical moment tensor of the mixture readily provides such an approximation.

1.2. Related works

Comparision to recent algorithms based on lattice basis reduction Two independent works (also independent and concurrent with our work) obtain polynomial-time algorithms for learning parallel-pancakes mixtures for the case that the component variance is zero along the hidden direction (infinitesimally flat pancakes)[6] Zadik et al. (2022); Diakonikolas and Kane (2022). These algorithms are based on the LLL lattice basis reduction algorithm Lenstra et al. (1982) and have a completely different flavor than our algorithms and previous algorithms for Gaussian mixture models. However, these lattice basis reduction techniques are expected to be brittle and limited to the case that the variance in the hidden direction is tiny.

Comparision to previous algorithms for mixtures with few components and unknown covariances Like our algorithms, many recent algorithms for learning GMMs make use of the sum-of-squares semidefinite programming hierarchy. While these algorithms and analyses have not been designed for our setting, we find it still instructive to discuss the differences and similarities to our algorithms.

Many of these algorithms also have in common that they employ the proof-to-algorithm paradigm, which has become the predominant way to analyze algorithms based on sum-of-squares for statistical estimation problems. (For expositions of this paradigm, see Barak and Steurer (2014); Raghavendra et al. (2018); Fleming et al. (2019).) This paradigm allows us to derive efficient estimation algorithms in a black-box way from identifiability proofs formalized in the sum-of-squares proof system.

As mentioned earlier, several recent algorithms consider mixtures of few well-separated Gaussian components with unknown covariances in the presence of adversarial outliers Bakshi et al. (2020a); Bakshi and Kothari (2020); Diakonikolas et al. (2020). While these algorithms have running times (at least) exponential in the number of components, their separation requirements are also exponentially stronger than ours. Even in the case that all covariances are the same (Σ) and the well-separatedness stems purely from the means $\mu_1, \ldots, \mu_k$, their identifiability proof requires separation $\|\Sigma^{-1/2}(\mu_a - \mu_b)\|^2 \geq k^{O(1)}$ (e.g., (Bakshi and Kothari, 2020, Lemma 4.16)). In constrast, our separation condition is logarithmic in k, which is the weakest separation condition, up to constant factors, that guarantees clusterability.

In order to deal with the kind of mild separation considered in this work, one could use one of the (robust) algorithms for parameter learning or density estimation of general k-component GMMs Moitra and Valiant (2010); Belkin and Sinha (2010); Bakshi et al. (2020b); Liu and Moitra

[6] These works also crucially assume a mild bound on the bit complexity of the unknown means.

(2021). While some of these works use separation between components in order to compute a rough partial clustering of far-away components as a pre-processing step, there doesn't appear to be a way to further exploit milder kinds of separation. For example, Moitra and Valiant (2010) learns the means of the mixture up to small error after projecting along a randomly chosen direction. This kind of projection cannot be expected to preserve any kind of separation of the high-dimensional mixture and even the sample complexity for recovering the means of this 1-dimensional mixture may be exponential in k (as shown in Moitra and Valiant (2010)). Both of the more recent works Bakshi et al. (2020b); Liu and Moitra (2021) end up enumerating subspaces related to the unknown parameters of the mixture. To the best of our knowledge, their approaches cannot avoid this step even for the kind of mildly-separated mixtures with lower bounded mixing weights considered in our work.

2. Techniques

We consider uniform[7] mixtures of $k \geq 2$ well-separated Gaussian components $N(\mu_1, \Sigma)$, ..., $N(\mu_k, \Sigma)$ with unknown means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and unknown covariance $\Sigma \in \mathbb{R}^{d \times d}$ (identical for all components). Here, we say components are *well-separated*[8] if the maximum affinity[9] (also called overlap) between two distinct components is bounded by $1/k^C$ for a large enough constant $C \geq 1$. For Gaussian components, this notion of well-separatedness means

$$\min_{a \neq b} \left\| \Sigma^{-1/2}(\mu_a - \mu_b) \right\| \gg \sqrt{\log k}\,. \tag{1}$$

Distinguishing well-separated mixtures from (pure) Gaussians Our algorithms are informed by investigating the parallel pancakes construction underlying Statistical Query lower bounds for such mixtures. This construction provides a mixture of k well-separated Gaussian components that appears to be exponentially hard to distinguish from the standard Gaussian distribution $N(0, I_d)$. In particular, this mixture matches the first $\Omega(k)$ moments of $N(0, 1)$.

The starting point of our algorithms is the following observation: In order for a mixture with $k \geq 2$ well-separated Gaussian components to match the first t moments of $N(0, I_d)$, the minimum mixing weight is necessarily smaller than $2^{-\Omega(t)}$. In particular, if the mixture has uniform mixing weights $\frac{1}{k}$, then always one of its first $O(\log k)$ moments distinguishes it from a standard Gaussian.

Underlying this observation is the following simple fact: *A distribution uniform over k real values can match no more than $O(\log k)$ moments of $N(0, 1)$.* To verify this fact, let $\boldsymbol{A}$ be a random variable uniformly distributed over k (not necessarily distinct) real values. Then, for all even integers $s \leq t$, the ratio of the normalized order-s and order-t moments of $\boldsymbol{A}$ is sandwiched in the following way,

$$k^{-1/s} \leq \frac{(\mathbb{E}\, \boldsymbol{A}^s)^{1/s}}{(\mathbb{E}\, \boldsymbol{A}^t)^{1/t}} \leq 1\,. \tag{2}$$

(This ratio is maximized if $\boldsymbol{A}$ is constant and minimized if $\mathbb{P}\{\boldsymbol{A} \neq 0\} = 1/k$.) In particular for $s = \log_2 k$, this ratio is lower bounded by $1/2$. On the other hand, for $\boldsymbol{B} \sim N(0, 1)$, the normalized

[7] In this section we restrict ourselves for ease of explanation to uniform mixtures. Our technical sections state all results for non-uniform mixtures.

[8] The term ”clusterable mixture” is sometimes used in the literature to refer to mixtures with well-separated components.

[9] The affinity of two probability measures is defined to be 1 minus their statistical distance Pollard (2002).

moments satisfy $(\mathbb{E}\,\boldsymbol{B}^r)^{1/r} = \Theta(r)^{1/2}$ and thus the ratio of normalized order-s and order-t moments is $\Theta(s/t)^{1/2}$. In particular, for some choice $t = \Theta(s)$, this ratio is smaller than $1/2$. It follows that for this choice of s and t, the ratios of normalized moments differ for $\boldsymbol{A}$ and for $\boldsymbol{B}$, which means that either their order-s or their order-t moments differ.[10]

This observation about uniform mixtures of $k \geq 2$ well-separated Gaussian components raises two questions: (1) do the first $O(\log k)$ moments also allow us to identify parameters of the mixture that are useful for clustering (in addition to allowing us to distinguish the mixture from $N(0,1)$), and (2) can we make make these results computationally efficient?

At a high level, we address question (1) by investigating ratios akin to Equation 2 between multivariate polynomials of degree $\Theta(\log k)$ derived from moments of the underlying mixture. To address question (2), we employ the proofs-to-algorithms paradigm (cf. Barak and Steurer (2014); Raghavendra et al. (2018)) and translate our arguments to syntactic proofs captured by the sum-of-squares proof system. These proofs then allow us to derive efficient algorithms (with running time $(kd)^{O(\log k)}$ or $(kd)^{(\log k)^{O(1)}}$) in a black-box way.

From decision to search: separating directions and ratios of moments In order to address question (1), we consider the goal of finding a direction $v \in \mathbb{R}^d$ that may be useful for clustering in the sense that along direction v, two of the components are significantly further apart than their standard deviation in this direction. More formally, we say that v is a *separating direction* for a mixture of k Gaussian components with unknown means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and unknown covariance $\Sigma \in \mathbb{R}^{d \times d}$ if there exist two means μ_a and μ_b such that

$$|\langle \mu_a - \mu_b, v\rangle| \gg \sqrt{\log k} \cdot \|\Sigma^{1/2} v\| \,. \tag{3}$$

We note that this direction v witnesses that the overlap of the components $N(\mu_a, \Sigma)$ and $N(\mu_b, \Sigma)$ is $k^{-\omega(1)}$. Conversely, whenever the overlap of two components is that small, there exists a vector v as above.

We aim to identify separating directions as solutions to inequalities between the following kind of moment polynomials: For $r \in \mathbb{N}$, we denote the degree-$2r$ *moment polynomial* $p_{2r} \in \mathbb{R}_{2r}[v]$ by

$$p_{2r}(v) := \mathbb{E}\langle \boldsymbol{y} - \boldsymbol{y}', v\rangle^{2r} \,, \tag{4}$$

where $\boldsymbol{y}, \boldsymbol{y}'$ are two independent random vectors identically distributed according to a uniform mixture of k Gaussian components $N(\mu_1, \Sigma), \ldots, N(\mu_k, \Sigma)$.

Using the fact that $\boldsymbol{y} - \boldsymbol{y}'$ can be expressed as a sum of two independent random vectors, one distributed uniformly over $\{\mu_a - \mu_b\}_{a,b\in[k]}$ and one distributed according to $N(0, 2\Sigma)$, these polynomials turn out to admit the following kind of approximation,

$$p_{2r}(v) = \left(k^{-2/r} \cdot \|Mv\|_{2r}^2 + \Theta(r) \cdot \|Av\|_2^2\right)^r \,. \tag{5}$$

Here, $A = \sqrt{2} \cdot \Sigma^{1/2}$, $M \in \mathbb{R}^{k^2 \times d}$ consists of the differences of means $(\mu_a - \mu_b)_{a,b\in[k]} \subseteq \mathbb{R}^d$ as rows and $\Theta(r)$ hides a nonnegative function upper bounded $O(r)$ and lower bounded by $\Omega(r/k^{2/r})$. (Since we will only consider $r \geq \log k$, we have $k^{-2/r} \geq \Omega(1)$.) Note that the first

[10] This proof shows that in order to distinguish a uniform distribution over k values from $N(0,1)$, it is enough to compare two moments of order logarithmic in k where the choice of orders depends only on k but not on the particular distribution.

term $k^{-2} \cdot \|Mv\|_{2r}^{2r}$ in (the binomial expansion of) Equation 5 corresponds to the order-r moment of the uniform distribution over $\{\mu_a - \mu_b\}_{a,b\in[k]}$ and the last term $\Theta(r)^r \cdot \langle v, 2\Sigma v\rangle^r$ to the order-r moment of $N(0, 2\Sigma)$.

We claim that for an appropriate choice $s \leq t$ with $s = \Theta(t) = \Theta(\log k)$, a direction v is separating in the sense of Equation 3 if and only $p_{2s}(v)^{1/s} \gtrsim p_{2t}(v)^{1/t}$. (Note that by convexity, $p_{2s}(v)^{1/s} \leq p_{2t}(v)^{1/t}$ holds for all directions v.) Underlying this claim is the familiar fact that for all $r \geq \log k$, the norm $\|Mv\|_{2r}$ equals up to constant factors the maximum entry of Mv, i.e., $\max_{a\neq b}|\langle \mu_a - \mu_b, v\rangle|$.

Indeed, suppose that v is a separating direction. Then, $p_{2s}(v)$ satisfies the lower bound,

$$p_{2s}(v)^{1/s} \geq k^{-2/s} \cdot \max_{a\neq b}\langle \mu_a - \mu_b, v\rangle^2 \,. \tag{6}$$

Since $s = \Theta(\log k)$, we have $p_{2s}(v)^{1/s} \gtrsim \max_{a\neq b}\langle \mu_a - \mu_b, v\rangle^2$. At the same time, $p_{2t}(v)$ satisfies the upper bound,

$$p_{2t}(v)^{1/t} \leq \max_{a\neq b}\langle \mu_a - \mu_b, v\rangle^2 + O(t) \cdot \|Av\|_2^2 \,. \tag{7}$$

Since v is a separating direction and $t = \Theta(\log k)$, the upper bound is dominated by the first term $\max_{a\neq b}\langle \mu_a - \mu_b, v\rangle^2$. Taking together both bounds, it follows that $p_{2s}(v)^{1/s} \gtrsim p_{2t}(v)^{1/t}$ for every separating direction v.

Conversely, suppose that $p_{2s}(v)^{1/s} \gtrsim p_{2t}(v)^{1/t}$ and our goal is to show that v is a separating direction. We lower bound $p_{2t}(v)$ using the last term in the approximation Equation 5 and apply the upper bound from Equation 7 to $p_{2s}(v)$. In this way, we obtain the inequality

$$\Omega(t) \cdot \|Av\|_2^2 \leq \max_{a\neq b}\langle \mu_a - \mu_b, v\rangle^2 + O(s) \cdot \|Av\|_2^2 \,. \tag{8}$$

By choosing t to be a large enough constant multiplied by s, we can ensure that the second term on the right-hand side is negligible. In this case, v satisfies $\max_{a\neq b}\langle \mu_a - \mu_b, v\rangle^2 \geq \Omega(t) \cdot \|Av\|_2^2$ for $t = \Theta(\log k)$, which means that v is a separating direction.

Challenges toward efficient algorithms for clustering Disregarding computational efficiency, the above characterization of separating directions in terms of ratios of moment polynomials suggests the following simple strategy for clustering uniform mixtures of Gaussian components $N(\mu_1, \Sigma)$, . . ., $N(\mu_k, \Sigma)$: we find an ϵ-cover of all separating directions by brute-force searching for an ϵ-cover of all solutions to an explicit polynomial system of the form $\left\{p_{2s}(v) = 1,\ p_{2t}(v) \leq O(1)^t\right\}$. Each separating direction gives us some information about what pairs of sample points belong to different components. For large enough mean separation $\min_{a\neq b}\|\Sigma^{-1/2}(\mu_a - \mu_b)\| \gg \sqrt{\log k}$, we can hope that by considering all such directions, we collect enough information to be able to extract a clustering of the sample that is approximately consistent with the components of the mixture.

This naive approach would require access only to moments of order $O(\log k)$ (which could be accurately estimated from a sample of size $d^{O(\log k)}$) but the running time is exponentially large (due to brute-force searching for solutions to a polynomial system).

A natural strategy to make this approach computationally efficient is the sum-of-squares hierarchy of semidefinite programming relaxations for systems of polynomial inequalities. Indeed, we can show that the above characterization of separating directions is faithfully captured by the sum-of-squares proof system underlying the sum-of-squares hierarchy. Unfortunately, it appears to be challenging to carry out the rounding step in full generality, i.e., extracting from the sum-of-squares hierarchy enough separating directions to separate all pairs of components and obtain a complete clustering of the sample.[11]

However, we can show that using the sum-of-squares hierarchy, it is possible to separate at least *some* pairs of components of the mixture by what we call a separating polynomial. Furthermore, for the special case of well-separated components with colinear means, we provide a more careful analysis and show that in this case the sum-of-squares hierarchy does offer enough information to extract a complete clustering.

Efficiently computing a separating polynomial As discussed above, we consider the goal of separating some pairs of components of a mixture (as opposed to the stronger goal of separating all pairs of components as would be required for a complete clustering). One way to achieve this goal is by finding a separating direction in the sense of Equation 3. In light of our previous characterization of separating directions, a natural starting point is a sum-of-squares relaxation for a polynomial system $\mathcal{A}$ of the form $\{p_{2s}(v) = 1,\ p_{2t}(v) \leq O(1)^t\}$ for appropriate $s \leq t$ satisfying $s = \Theta(t) = \Theta(\log k)$.

Unfortunately, the structure of the set of separating directions does not appear to be amenable to the usual kind of rounding techniques for sum-of-squares relaxations, and it appears to be challenging to extract a single separating direction. To overcome this obstacle, we allow our rounding procedure to output a more general object, called a *separating polynomial*, that still allows us to separate some pairs of mixture components.

Recall that a solution to a sum-of-squares relaxation for a polynomial system $\mathcal{A}$ can be interpreted as *pseudo-distribution* D that behaves in certain ways like a distribution supported on vectors satisfying $\mathcal{A}$. More concretely, the pseudo-distribution D satsifies (in expectation) all polynomial inequalities that can be derived syntactically from $\mathcal{A}$ by a low-degree sum-of-squares proof (see Section A, especially Definition 4). The previously discussed characterization of separating directions in terms of the polynomial system $\mathcal{A}$ turns out to be captured by low-degree sum-of-squares proofs. Concretely, we can derive from $\mathcal{A}$ via low-degree sum-of-squares proof the polynomial inequality[12] $\|Mv\|_{2s}^{2s} \geq (C\log k)^s \cdot \|Av\|_2^{2s}$ (corresponding to Equation 3). Here, $C \geq 1$ is an absolute constant that we can choose as large as we like. Consequently, the pseudo-distribution D satisfies this inequality in expectation, $\tilde{\mathbb{E}}_{D(v)}\|Mv\|_{2s}^{2s} \geq (C\log k)^s \cdot \tilde{\mathbb{E}}_{D(v)}\|Av\|_2^{2s}$. By linearity of (pseudo-)expectation, there exist distinct components $a \neq b$ such that

$$\tilde{\mathbb{E}}_{D(v)} \langle \mu_a - \mu_b, v\rangle^{2s} \geq k^{-2} \tilde{\mathbb{E}}_{D(v)} \|Mv\|_{2s}^{2s} \geq (k^{-2/s} \cdot C\log k)^s \cdot \tilde{\mathbb{E}}_{D(v)} \|Av\|_2^{2s} \,. \tag{9}$$

We extract the following polynomial from this pseudo-distribution,

$$q(u) := \tilde{\mathbb{E}}_{D(v)} \langle u, v\rangle^{2s} \,. \tag{10}$$

[11] In the context of estimation problems, rounding procedures for sum-of-squares hierarchies tend to work well if there is a unique target solution (e.g., a planted sparse vector in a random subspace) or if there is a small number of target solutions (e.g., the components of a low-rank tensor). One could try to simplify the structure of the set of separating directions, e.g., by focusing on ”extreme” separating directions of the form $v = \Sigma^{-1}(\mu_a - \mu_b)$. Unfortunately, we do not know the same kind of characterization in terms of polynomial inequalities for such a simplified set of separating directions.

[12] Here, we reuse the notation introduced in the context of Equation 5.

By construction, $q(\mu_a - \mu_b)$ equals the left-hand side of Equation 9. At the same time, letting $\boldsymbol{y}, \boldsymbol{y}' \sim N(\mu_c, \Sigma)$ and $\boldsymbol{w} \sim N(0, I_d)$, we have

$$\begin{aligned} \mathbb{E}\, q(\boldsymbol{y} - \boldsymbol{y}') &= \mathbb{E}\, q(A\boldsymbol{w}) \\ &= \underset{D(v)}{\tilde{\mathbb{E}}}\, \mathbb{E} \langle v, A\boldsymbol{w} \rangle^{2s} \\ &= (2s-1)!! \cdot \underset{D(v)}{\tilde{\mathbb{E}}} \|A\boldsymbol{w}\|_2^{2s} \end{aligned}$$

Consequently, since $s = \Theta(\log k)$ and $(2s-1)!! \leq O(s)^s$,

$$\frac{q(\mu_a - \mu_b)}{\mathbb{E}\, q(A\boldsymbol{w})} \geq \frac{(k^{-2/s} \cdot C \log k)^s}{(2s-1)!!} \geq \Omega(C)^s \,. \tag{11}$$

For an appropriate choice of $C \geq 1$, the right-hand side above is at least 10^s. Since by convexity $\mathbb{E}\, q(\mu_a - \mu_b + A\boldsymbol{w}) \geq q(\mu_a - \mu_b)$, we obtain the following inequality,

$$\frac{\mathbb{E}\, q(\mu_a - \mu_b + A\boldsymbol{w})}{\mathbb{E}\, q(A\boldsymbol{w})} \geq 10^s \,. \tag{12}$$

This inequality shows that the polynomial $q(u)$ separates the components $N(\mu_a, \Sigma)$ and $N(\mu_b, \Sigma)$ in the following sense: The numerator of Equation 12 is the typical value of $q(\boldsymbol{y} - \boldsymbol{y}')$ for $\boldsymbol{y} \sim N(\mu_a, \Sigma)$ and $\boldsymbol{y}' \sim N(\mu_b, \Sigma)$. The denominator of Equation 12 is the typical value of $q(\boldsymbol{y} - \boldsymbol{y}')$ for $\boldsymbol{y}, \boldsymbol{y}' \sim N(\mu_c, \Sigma)$ and all $c \in [k]$. Equation 12 asserts that the gap between these values is at least 10^s.

The polynomial $q(u)$ can be used to compute a bipartition of the sample that separates at least one pair of components. Note that $q(u)^{1/2s} = (\tilde{\mathbb{E}}_{D(v)} \langle u, v \rangle^{2s})^{1/2s}$ satisfies the triangle inequality (see Lemma 4.5 in Barak and Steurer (2014)). Then we can define the distance function $d_q(x, y) = q(x - y)^{1/2s}$ and use it in a greedy algorithm in order to obtain the bipartition.

Efficiently computing a clustering for colinear means For the case that the means are colinear, we consider a strengthening of our previous approach. Instead of trying to solve a polynomial system of the form $\{p_{2s}(v) = 1,\ p_{2t}(v) \leq O(1)^t\}$, we aim to solve the following related optimization problem:

$$\text{minimize} \quad \frac{p_{2t}(v)^{1/t}}{p_{2s}(v)^{1/s}} \quad \text{subject to} \quad v \in \mathbb{R}^d. \tag{13}$$

Algorithmically, we again employ an appropriate sum-of-squares formulation.

To simplify some of our arguments, it is useful to preprocess the mixture and bring $\boldsymbol{y} - \boldsymbol{y}'$ in isotropic position so that $\frac{1}{k^2} \sum_{a,b=1}^k (\mu_a - \mu_b)(\mu_a - \mu_b)^\mathsf{T} + 2\Sigma = I_d$. (Here, $\boldsymbol{y}, \boldsymbol{y}'$ are two independent random vectors distributed according to the mixture.) For every vector v, we denote by $v^\parallel$ its orthogonal projection into the span of $\{\mu_a - \mu_b\}_{a,b \in [k]}$ and by $v^\perp = v - v^\parallel$ its projection into the orthogonal complement.

Every optimizer v of Equation 13 necessarily satisfies,

$$\frac{p_{2t}(v)^{1/t}}{p_{2s}(v)^{1/s}} \leq \frac{p_{2t}(v^\parallel)^{1/t}}{p_{2s}(v^\parallel)^{1/s}} \leq O(1) \,. \tag{14}$$

Here, the upper bound $O(1)$ hides an absolute constant whenever we have well-separated components and $\log k \leq s \leq t$. The argument for this upper bound is similar to our discussion for the characterization of separating directions.

We can also use the decomposition $v = v^{\|} + v^{\perp}$ for our previous approximation Equation 5 of moment polynomials,

$$p_{2r}(v) = \left(k^{-2/r} \cdot \left\|Mv^{\|}\right\|_{2r}^{2} + \Theta(r) \cdot \left(\left\|Av^{\|}\right\|_{2}^{2} + \left\|v^{\perp}\right\|_{2}^{2}\right)\right)^{r} . \tag{15}$$

Here, we use that after bringing $\boldsymbol{y} - \boldsymbol{y}'$ in isotropic position, the covariance Σ acts as identity orthogonal to the span of $\{\mu_a - \mu_b\}_{a,b\in[k]}$. In particular, $Av = Av^{\|} + v^{\perp}$ and $\|Av\|_2^2 = \|Av^{\|}\|_2^2 + \|v^{\perp}\|_2^2$.

An immediate consequence of Equation 15 is the following representation of the ratio we seek to minimize,

$$\frac{p_{2t}(v)^{1/t}}{p_{2s}(v)^{1/s}} = \frac{p_{2t}(v^{\|})^{1/t} + \Theta(t) \cdot \|v^{\perp}\|_2^2 \pm \Theta(t) \cdot \|Av^{\|}\|_2^2}{p_{2s}(v^{\|})^{1/s} + \Theta(s) \cdot \|v^{\perp}\|_2^2 \pm \Theta(s) \cdot \|Av^{\|}\|_2^2} . \tag{16}$$

We claim that for an appropriate choice of s and t Equation 16 and Equation 14 together imply that $\|v^{\perp}\| \lesssim \|Av^{\|}\|$. Indeed, for the sake of a contradiction, suppose $\|v^{\perp}\| \gg \|Av^{\|}\|$ so that the terms involving $\|Av^{\|}\|$ in Equation 16 are negligible. But then, if we choose t as s times a sufficiently larger constant factor, the remaining ratio $\frac{p_{2t}(v^{\|})^{1/t} + \Theta(t)\cdot\|v^{\perp}\|_2^2}{p_{2s}(v^{\|})^{1/s} + \Theta(s)\cdot\|v^{\perp}\|_2^2}$ is strictly bigger than $\frac{p_{2t}(v^{\|})^{1/t}}{p_{2s}(v^{\|})^{1/s}}$, which contradicts our optimality condition Equation 14. (For this argument, we are also using the previous upper bound $\frac{p_{2t}(v^{\|})^{1/t}}{p_{2s}(v^{\|})^{1/s}} \leq O(1)$ from Equation 14.)

It turns out that in order to compute a clustering for colinear means, it suffices to find a vector satisfying $\|v^{\perp}\| \lesssim \|Av^{\|}\|$.

We note that the algorithm we present in Section C and Section D to find such a direction v follows a somewhat different strategy and minimizes ratios of the form $\|v\|_2^2/p_{2s}(v)^{1/s}$ or $p_{2t}(v)^{1/t}/\|v\|_2^2$ via appropriate sum-of-squares formulations.

Acknowledgments

The project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 815464). We thank Samuel B. Hopkins for discussions related to this project.

References

Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *J. ACM*, 67(6):32:1–32:42, 2020. doi: 10.1145/3417994. URL https://doi.org/10.1145/3417994.

Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *CoRR*, abs/2005.02970, 2020. URL https://arxiv.org/abs/2005.02970.

Ainesh Bakshi, Ilias Diakonikolas, Samuel B. Hopkins, Daniel Kane, Sushrut Karmalkar, and Pravesh K. Kothari. Outlier-robust clustering of Gaussians and other non-spherical mixtures.

In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 149–159. IEEE Computer Soc., Los Alamitos, CA, 2020a. doi: 10.1109/FOCS46700.2020.00023. URL https://doi.org/10.1109/FOCS46700.2020.00023.

Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. Robustly learning mixtures of k arbitrary gaussians. *CoRR*, abs/2012.02119, 2020b. URL https://arxiv.org/abs/2012.02119.

Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. In *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. IV*, pages 509–533. Kyung Moon Sa, Seoul, 2014.

Mikhail Belkin and Kaushik Sinha. Toward learning gaussian mixtures with arbitrary separation. In *COLT*, pages 407–419. Omnipress, 2010.

S. Charles Brubaker and Santosh Vempala. Isotropic PCA and affine-invariant clustering. In *FOCS*, pages 551–560. IEEE Computer Society, 2008.

Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous lwe. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 694–707, 2021.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *FOCS*, pages 634–644. IEEE Computer Society, 1999.

Ilias Diakonikolas and Daniel Kane. Non-gaussian component analysis via lattice basis reduction. In *Conference on Learning Theory*, pages 4535–4547. PMLR, 2022.

Ilias Diakonikolas and Daniel M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 184–195. IEEE Computer Soc., Los Alamitos, CA, 2020.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures (extended abstract). In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 73–84. IEEE Computer Soc., Los Alamitos, CA, 2017. doi: 10.1109/FOCS.2017.16. URL https://doi.org/10.1109/FOCS.2017.16.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060. ACM, New York, 2018. doi: 10.1145/3188745.3188758. URL https://doi.org/10.1145/3188745.3188758.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.*, 48(2):742–864, 2019. ISSN 0097-5397. doi: 10.1137/17M1126680. URL https://doi.org/10.1137/17M1126680.

Ilias Diakonikolas, Samuel B. Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *CoRR*, abs/2005.06417, 2020. URL https://arxiv.org/abs/2005.06417.

Noah Fleming, Pravesh Kothari, and Toniann Pitassi. Semialgebraic proofs and efficient algorithm design. *Found. Trends Theor. Comput. Sci.*, 14(1-2):1–221, 2019. doi: 10.1561/0400000086. URL https://doi.org/10.1561/0400000086.

Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of gaussians in high dimensions. In *STOC*, pages 761–770. ACM, 2015.

Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous lwe is as hard as lwe & applications to learning gaussian mixtures. *arXiv preprint arXiv:2204.02550*, 2022.

Samuel B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. ACM, New York, 2018. doi: 10.1145/3188745.3188748. URL https://doi.org/10.1145/3188745.3188748.

Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS'13—Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science*, pages 11–19. ACM, New York, 2013.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *STOC*, pages 553–562. ACM, 2010.

Pravesh K. Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017.

Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, New York, 2018.

A. K. Lenstra, H. W. Lenstra, Jr., and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261(4):515–534, 1982. ISSN 0025-5831. doi: 10.1007/BF01457454. URL http://dx.doi.org/10.1007/BF01457454.

Allen Liu and Jerry Li. Clustering mixtures with almost optimal separation in polynomial time. In *STOC '22—Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261. ACM, New York, 2022.

Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaussians. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 518–531. ACM, 2021. doi: 10.1145/3406325.3451084. URL https://doi.org/10.1145/3406325.3451084.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS*, pages 93–102. IEEE Computer Society, 2010.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. ISSN 02643820. URL http://www.jstor.org/stable/90667.

David Pollard. *A user's guide to measure theoretic probability*. Number 8. Cambridge University Press, 2002.

Prasad Raghavendra, Tselil Schramm, and David Steurer. High dimensional estimation via sum-of-squares proofs. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. IV. Invited lectures*, pages 3389–3423. World Sci. Publ., Hackensack, NJ, 2018.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixtures of distributions. In *FOCS*, page 113. IEEE Computer Society, 2002.

Ilias Zadik, Min Jae Song, Alexander S Wein, and Joan Bruna. Lattice-based methods surpass sum-of-squares in clustering. In *Conference on Learning Theory*, pages 1247–1248. PMLR, 2022.

Appendix A. Preliminaries

In this section we introduce sum-of-squares proofs and their duals, pseudo-distributions and pseudo-expectations.

Sum-of-squares proofs.

Definition 4 (Sum-of-squares proofs) *Let $p(x)$ and $q_1(x), ..., q_m(x)$ be polynomials over $x \in \mathbb{R}^n$ and let $\mathcal{A} = \{q_1(x) \geq 0, ..., q_m(x) \geq 0\}$ be a system of polynomial inequalities. A sum-of-squares proof of degree t that $p(x) \geq 0$ under $\mathcal{A}$ is an identity of the form*

$$p(x) = \sum_{S \subseteq [m]} \left(\sum_{i=1}^{m_S} r_{S,i}(x)^2 \right) \prod_{j \in S} q_j(x) \tag{17}$$

for polynomials $r_{S,i}(x)$, such that $\max_{S,i} \deg(r_{S,i}(x)^2 \prod_{j \in S} q_j(x)) \leq t$.

If there exists a sum-of-squares proof of degree t that $p(x) \geq 0$ under $\mathcal{A}$, we write $\mathcal{A} \vdash_{t}^{x} p(x) \geq 0$. We also use the notation $\mathcal{A} \vdash_{t}^{x} p(x) \geq q(x)$ if $\mathcal{A} \vdash_{t}^{x} p(x) - q(x) \geq 0$ and $\mathcal{A} \vdash_{t}^{x} p(x) \leq q(x)$ if $\mathcal{A} \vdash_{t}^{x} q(x) - p(x) \geq 0$. If $\mathcal{A} = \emptyset$, we omit it altogether and write $\vdash_{t}^{x} p(x) \geq 0$. We also sometimes omit $\mathcal{A}$ if it is clear from context what axioms are assumed. We note that if $\mathcal{A} \vdash_{t}^{x} p(x) \geq q(x)$ and $\mathcal{A} \vdash_{t}^{s} q(x) \geq r(x)$, then $\mathcal{A} \vdash_{t}^{x} p(x) \geq r(x)$, which allows writing chains of inequalities of the form $\mathcal{A} \vdash_{t}^{x} p(x) \geq s(x) \geq r(x)$.

Pseudo-distributions and pseudo-expectations. We begin by defining pseudo-distributions and pseudo-expectations.

Definition 5 (Pseudo-distributions) *A pseudo-distribution D of degree t is a function from $\mathbb{R}^n$ to $\mathbb{R}$ with finite support such that $\sum_{x \in \mathrm{supp}(D)} D(x) = 1$ and $\sum_{x \in \mathrm{supp}(D)} D(x)p(x)^2 \geq 0$ for all polynomials $p(x)$ with $\deg(p(x)^2) \leq t$.*

Definition 6 (Pseudo-expectations) *Given a pseudo-distribution D of degree t, the associated pseudo-expectation $\tilde{\mathbb{E}}_{D(x)}$ is defined by $\tilde{\mathbb{E}}_{D(x)} f(x) = \sum_{x \in \mathrm{supp}(D)} D(x) f(x)$ for a function $f(x)$.*

We now define the notion of a pseudo-distribution that satisfies a set of polynomial inequalities.

Definition 7 (Constrained pseudo-distributions) *A pseudo-distribution D of degree t satisfies the set of polynomial inequalities $\mathcal{A} = \{q_1(x) \geq 0, ..., q_m(x) \geq 0\}$ if, for all $S \subseteq [m]$, it holds that $\tilde{\mathbb{E}}_{D(x)} r(x)^2 \prod_{j \in S} q_j(x) \geq 0$ for all polynomials $r(x)$ such that $\deg(r(x)^2 \prod_{j \in S} q_j(x)) \leq t$.*

D approximately satisfies $\mathcal{A}$ up to error η if, under the same conditions as in the previous case, $\tilde{\mathbb{E}}_{D(x)} r(x)^2 \prod_{j \in S} q_j(x) \geq -\eta \|r(x)^2\|_2 \prod_{j \in S} \|q_j(x)\|_2$, where $\|p(x)\|_2$ denotes the 2-norm of the vector of coefficients of the polynomial $p(x)$.

The connection between pseudo-distributions and sum-of-squares proofs is made in Fact 8, which shows that if a pseudo-distribution satisfies a set of polynomial inequalities, then it also satisfies any other polynomial inequalities derived from this set through sum-of-squares proofs.

Fact 8 *If D is a pseudo-distribution of degree t that satisfies $\mathcal{A}$ and if $\mathcal{A} \,\Big|\frac{x}{s}\, p(x) \geq 0$, then $\tilde{\mathbb{E}}_{D(x)} r(x)^2 p(x) \geq 0$ for all polynomials $r(x)$ such that $\deg(r(x)^2 p(x)) \leq t$. If D approximately satisfies $\mathcal{A}$ up to error η, then, under the same conditions as in the previous case, $\tilde{\mathbb{E}}_{D(x)} r(x)^2 p(x) \geq -\eta \|r(x)^2\|_2 \|p(x)\|_2$.*

Finally, Fact 9 shows that there exists an algorithm with time complexity $(n+m)^{O(t)}$ to compute a pseudo-distribution of degree t that approximately satisfies $\mathcal{A}$ up to error $2^{-n^{\Theta(t)}}$.

Fact 9 *For $x \in \mathbb{R}^n$, if $\mathcal{A} = \{q_1(x) \geq 0, ..., q_m(x) \geq 0\}$ is feasible and explicitly bounded [13], then there exists an algorithm that runs in time $(n+m)^{O(t)}$ and computes a pseudo-distribution of degree t that approximately satisfies $\mathcal{A}$ up to error $2^{-n^{\Theta(t)}}$ [14].*

Appendix B. Separating polynomial

Setting. We consider a mixture of k Gaussian distributions $N(\mu_i, \Sigma)$ with mixing weights p_i for $i = 1, ..., k$, where $\mu_i \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite, and $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$. Let $p_{\min} = \min_i p_i$.

The distribution satisfies mean separation for at least one pair of means: for some $C_{sep} > 0$, there exist $a, b \in [k]$ such that

$$\left\| \Sigma^{-1/2}(\mu_a - \mu_b) \right\|^2 \geq C_{sep} \log p_{\min}^{-1}.$$

Theorem 10 (Separating polynomial algorithm) *Consider the Gaussian mixture model defined above, with C_{sep} larger than some universal constant. Let $n_0 = (p_{\min}^{-1} d)^{O(\log p_{\min}^{-1})}$. Given a sample of size $n \geq n_0$ from the mixture, there exists an algorithm that computes in time $n \cdot d^{O(\log p_{\min}^{-1})}$ a d-variate degree-$O(\log p_{\min}^{-1})$ polynomial q such that with high probability the following two properties hold. Let $s = \lceil \log p_{\min}^{-1} \rceil$. Then:*

- *There exist distinct $a, b \in [k]$ such that the independent random vectors $\boldsymbol{y} \sim N(\mu_a, \Sigma)$ and $\boldsymbol{y}' \sim N(\mu_b, \Sigma)$ satisfy*

$$\mathbb{P}\left\{ q(\boldsymbol{y} - \boldsymbol{y}') \geq \frac{1}{20^s} \right\} \geq 0.99999.$$

[13] Explicit boundedness means that $\mathcal{A}$ contains a constraint of the form $x_1^2 + ... + x_n^2 \leq B$. In our applications it is possible to add such a constraint with B large enough such that the constraint is always satisfied by the intended solution.

[14] In our applications this error is negligible.

- *For all $a \in [k]$, the independent random vectors $\boldsymbol{y}, \boldsymbol{y}' \sim N(\mu_a, \Sigma)$ satisfy*

$$\mathbb{P}\left\{q(\boldsymbol{y} - \boldsymbol{y}') \leq \frac{1}{200^s}\right\} \geq 0.99999.$$

Theorem 11 (Separating bipartition algorithm) *Consider the Gaussian mixture model defined above, with C_{sep} larger than some universal constant. Let $n_0 = (p_{\min}^{-1} d)^{O(\log p_{\min}^{-1})}$. Given a sample of size $n \geq n_0$ from the mixture, there exists an algorithm that runs in time $n \cdot d^{O(\log p_{\min}^{-1})}$ and returns with probability 0.99 a partition of $[n]$ into two sets C_1 and C_2 such that, if true clustering of the samples is $S_1, ..., S_k$, then*

$$\max_i \frac{|C_1 \cap S_i|}{|S_i|} \geq 0.99 \quad \textit{and} \quad \max_i \frac{|C_2 \cap S_i|}{|S_i|} \geq 0.99.$$

We introduce some further notation for this section. Let the random variable $\boldsymbol{z} \in \mathbb{R}^d$ be distributed according to the difference of two independent samples from the mixture. Then $\boldsymbol{z}$ is distributed according to a mixture of Gaussians $N(\mu_i - \mu_j, 2\Sigma)$ with mixing weights $p_i p_j$ for all $i, j \in [k]$. Let $\Sigma_z = 2\Sigma$, let $\boldsymbol{\mu}_z$ be $\mu_i - \mu_j$ with probability $p_i p_j$, and let $\boldsymbol{w}_z \sim N(0, \Sigma_z)$. Then we also have that $\boldsymbol{z} = \boldsymbol{\mu}_z + \boldsymbol{w}_z$, with $\boldsymbol{\mu}_z$ and $\boldsymbol{w}_z$ independent of each other.

B.1. Exact moment results

The main ingredient of the algorithm is Lemma 12, stated below. This lemma shows that, given a pseudo-expectation that satisfies the moment lower bound $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} \geq c^s$ and the moment upper bound $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq C^t$ for $s \ll t$, it is possible to construct a separating polynomial. Note that the constraints that the pseudo-expectation satisfies are expressed in terms of exact moments of the distribution, to which we do not have access. Finite sample considerations are discussed starting with Section B.2.

Lemma 12 (Separating polynomial from pseudo-expectation) *Let $c > 0$ and $C \geq 0$. Let $s \geq 1$ and $t \geq 50000Cs/c$ integers. Given a pseudo-expectation $\tilde{\mathbb{E}}$ of degree at least $2t$ over a variable $v \in \mathbb{R}^d$ that satisfies $\{\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} \geq c^s, \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq C^t\}$, let $q(u) = \langle \tilde{\mathbb{E}} v^{\otimes 2s}, u^{\otimes 2s}\rangle$. Then:*

- *There exist distinct $a, b \in [k]$ such that the independent random vectors $\boldsymbol{y} \sim N(\mu_a, \Sigma)$ and $\boldsymbol{y}' \sim N(\mu_b, \Sigma)$ satisfy*

$$\mathbb{P}\left\{q(\boldsymbol{y} - \boldsymbol{y}') \geq \frac{1}{2}\left(\frac{c}{16}\right)^s\right\} \geq 0.99999\,.$$

- *For all $a \in [k]$, the independent random vectors $\boldsymbol{y}, \boldsymbol{y}' \sim N(\mu_a, \Sigma)$ satisfy*

$$\mathbb{P}\left\{q(\boldsymbol{y} - \boldsymbol{y}') \leq 320\left(\frac{4Cs}{t}\right)^s\right\} \geq 0.99999\,.$$

In what follows, we prove a number of supporting lemmas, after which we prove Lemma 12. Then, we state and prove Lemma 17, which shows that there exists a vector $v \in \mathbb{R}^d$ which satisfies the constraints required by Lemma 12.

We proceed with the supporting lemmas. Lemma 13 and Lemma 14 give sum-of-squares bounds on the moments of the mixture.

Lemma 13 (Moment upper bound) *For $t \geq 1$ integer,*

$$\vdash^{v}_{2t} \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq 2^{2t-1}\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} + 2^{2t-1}(v^\top \Sigma_z v)^t t^t .$$

Proof

$$\begin{aligned}\vdash^{v}_{2t} \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} &= \mathbb{E}(\langle \boldsymbol{\mu}_z, v\rangle + \langle \boldsymbol{w}_z, v\rangle)^{2t} \overset{(1)}{\leq} 2^{2t-1}\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} + 2^{2t-1}\mathbb{E}\langle \boldsymbol{w}_z, v\rangle^{2t} \\ &\overset{(2)}{\leq} 2^{2t-1}\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} + 2^{2t-1}(v^\top \Sigma_z v)^t t^t\end{aligned}$$

where in (1) we used Lemma 52 and in (2) we used that $\mathbb{E}\langle \boldsymbol{w}_z, v\rangle^{2t} = (v^\top \Sigma_z v)^t (2t-1)!! \leq (v^\top \Sigma_z v)^t t^t$. ■

Lemma 14 (Moment lower bound) *For $t \geq 1$ integer,*

$$\vdash^{v}_{2t} \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \geq \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} + (v^\top \Sigma_z v)^t \frac{t^t}{2^t}.$$

Proof

$$\begin{aligned}\vdash^{v}_{2t} \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} &= \mathbb{E}(\langle \boldsymbol{\mu}_z, v\rangle + \langle \boldsymbol{w}_z, v\rangle)^{2t} = \sum_{j=0}^{2t} \binom{2t}{j} \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{j} \langle \boldsymbol{w}_z, v\rangle^{2t-j} \\ &\overset{(1)}{=} \sum_{j=0}^{2t} \binom{2t}{j} \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{j} \mathbb{E}\langle \boldsymbol{w}_z, v\rangle^{2t-j} \overset{(2)}{=} \sum_{s=0}^{t} \binom{2t}{2s} \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2s} \mathbb{E}\langle \boldsymbol{w}_z, v\rangle^{2t-2s} \\ &\geq \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} + \mathbb{E}\langle \boldsymbol{w}_z, v\rangle^{2t} \overset{(3)}{\geq} \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} + (v^\top \Sigma_z v)^t \frac{t^t}{2^t}\end{aligned}$$

where in (1) we used that $\boldsymbol{\mu}_z$ and $\boldsymbol{w}_z$ are independent, in (2) we used that $\mathbb{E}\langle \boldsymbol{w}_z, v\rangle^{2t-j} = 0$ for $2t - j$ odd, and in (3) we used that $\mathbb{E}\langle \boldsymbol{w}_z, v\rangle^{2t} = (v^\top \Sigma_z v)^t (2t-1)!! \geq (v^\top \Sigma_z v)^t \frac{t^t}{2^t}$. ■

Going forward, Lemma 15 proves that, if the moments of $\boldsymbol{z}$ are small in direction v, then the variance of the components of the mixture is also small in direction v. Given in addition an upper bound on the moments of $\boldsymbol{z}$ in direction v for a sufficiently large moment, Lemma 16 proves that the contribution of the means in direction v is large.

Lemma 15 *Let $C \geq 0$. For $t \geq 1$ integer,*

$$\left\{\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq C^t\right\} \vdash^{v}_{2t} \left\{v^\top \Sigma_z v \leq \frac{2C}{t}\right\}.$$

Proof Substitute the lower bound of Lemma 14 into the axiom:

$$\vdash^{v}_{2t} \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} + (v^\top \Sigma_z v)^t t^t / 2^t \leq C^t .$$

Use that $\vdash^{v}_{2t} \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} \geq 0$ to drop the first term and then divide by $t^t/2^t$. This proves that $\vdash^{v}_{2t} (v^\top \Sigma_z v)^t \leq 2^t C^t / t^t$. Finally, by Lemma 56, this implies that $\vdash^{v}_{2t} v^\top \Sigma_z v \leq 2C/t$. ■

Lemma 16 *Let $c > 0$ and $C \geq 0$. For $s \geq 1$ and $t \geq 16Cs/c$ integers,*

$$\{\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} \geq c^s, \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq C^t\} \vdash^{v}_{2t} \left\{\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2s} \geq \left(\frac{c}{4}\right)^s\right\}.$$

Proof Substitute the upper bound of Lemma 13 into the first axiom:

$$\vdash^{v}_{2s} 2^{2s-1}\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2s} + 2^{2s-1}(v^\top \Sigma_z v)^s s^s \geq c^s.$$

Use, by Lemma 15 and Lemma 55, that $\vdash^{v}_{2t} (v^\top \Sigma_z v)^s \leq (2C/t)^s$:

$$\vdash^{v}_{2s} 2^{2s-1}\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2s} + (8Cs/t)^s \geq c^s.$$

Then use that $t \geq 16Cs/c$ to obtain that $\vdash^{v}_{2s} 2^{2s-1}\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2s} \geq c^s/2$. Finally, divide by 2^{2s-1} to obtain that $\vdash^{v}_{2s} \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2s} \geq (c/4)^s$. ■

Now we prove Lemma 12.

Proof of Lemma 12 Let $\boldsymbol{w} \sim N(0, I_d)$. Note that, for $\boldsymbol{y} \sim N(\mu_a, \Sigma)$ and $\boldsymbol{y}' \sim N(\mu_b, \Sigma)$ we have that $\boldsymbol{y} - \boldsymbol{y}' \sim \mu_a - \mu_b + \Sigma_z^{1/2}\boldsymbol{w}$, so $q(\boldsymbol{y} - \boldsymbol{y}') = q(\mu_a - \mu_b + \Sigma_z^{1/2}\boldsymbol{w})$. Similarly, for $\boldsymbol{y}, \boldsymbol{y}' \sim N(\mu_a, \Sigma)$ we have that $\boldsymbol{y} - \boldsymbol{y}' \sim \Sigma_z^{1/2}\boldsymbol{w}$, so $q(\boldsymbol{y} - \boldsymbol{y}') = q(\Sigma_z^{1/2}\boldsymbol{w})$. Therefore, we want to show (1) that there exist distinct $a, b \in [k]$ such that $q(\mu_a - \mu_b + \Sigma_z^{1/2}\boldsymbol{w})$ is large and (2) that $q(\Sigma_z^{1/2}\boldsymbol{w})$ is small.

By Lemma 16, we have $\tilde{\mathbb{E}}\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2s} \geq (c/4)^s$. By linearity, we also have $\mathbb{E}\tilde{\mathbb{E}}\langle \boldsymbol{\mu}_z, v\rangle^{2s} \geq (c/4)^s$, so $\mathbb{E}q(\boldsymbol{\mu}_z) \geq (c/4)^s$. Therefore there exists some μ_z in the support of $\boldsymbol{\mu}_z$ such that $q(\mu_z) \geq (c/4)^s$. Therefore, there exist $a, b \in [k]$ such that $q(\mu_a - \mu_b) \geq (c/4)^s$. Furthermore, a and b are distinct, because otherwise $q(\mu_a - \mu_b) = 0$.

We attempt to lower bound $q(\mu_a - \mu_b + \Sigma_z^{1/2}\boldsymbol{w})$:

$$\begin{aligned}
q(\mu_a - \mu_b + \Sigma_z^{1/2}\boldsymbol{w}) &= \langle \tilde{\mathbb{E}}v^{\otimes 2s}, (\mu_a - \mu_b + \Sigma_z^{1/2}\boldsymbol{w})^{\otimes 2s}\rangle \\
&= \tilde{\mathbb{E}}\langle v, \mu_a - \mu_b + \Sigma_z^{1/2}\boldsymbol{w}\rangle^{2s} \\
&\geq \frac{1}{2^{2s-1}}\tilde{\mathbb{E}}\langle v, \mu_a - \mu_b\rangle^{2s} - \tilde{\mathbb{E}}\langle v, \Sigma_z^{1/2}\boldsymbol{w}\rangle^{2s} \\
&= \frac{1}{2^{2s-1}}q(\mu_a - \mu_b) - q(\Sigma_z^{1/2}\boldsymbol{w}) \\
&\geq (c/16)^s - q(\Sigma_z^{1/2}\boldsymbol{w}).
\end{aligned}$$

We want to show that $q(\Sigma_z^{1/2}\boldsymbol{w})$ is small with high probability. We start by analyzing the mean and second moment of $q(\Sigma_z^{1/2}\boldsymbol{w})$. For $\ell \in \{1, 2\}$, we have

$$\begin{aligned}
\mathbb{E}q(\Sigma_z^{1/2}\boldsymbol{w})^\ell &= \mathbb{E}\left(\langle \tilde{\mathbb{E}}v^{\otimes 2s}, (\Sigma_z^{1/2}\boldsymbol{w})^{\otimes 2s}\rangle\right)^\ell = \mathbb{E}\left(\tilde{\mathbb{E}}\langle v, \Sigma_z^{1/2}\boldsymbol{w}\rangle^{2s}\right)^\ell \\
&\overset{(1)}{\leq} \mathbb{E}\tilde{\mathbb{E}}\langle v, \Sigma_z^{1/2}\boldsymbol{w}\rangle^{2s\ell} = \tilde{\mathbb{E}}\mathbb{E}\langle v, \Sigma_z^{1/2}\boldsymbol{w}\rangle^{2s\ell} \\
&= \tilde{\mathbb{E}}\mathbb{E}\langle \Sigma_z^{1/2}v, \boldsymbol{w}\rangle^{2s\ell} \leq (s\ell)^{s\ell}\tilde{\mathbb{E}}(v^\top \Sigma_z v)^{s\ell} \\
&\overset{(2)}{\leq} (2CS\ell/t)^{s\ell},
\end{aligned}$$

where in (1) for $\ell = 2$ we used Lemma 61 and in (2) we used that, by Lemma 15 and Lemma 55, $\tilde{\mathbb{E}}(v^\top \Sigma_z v)^{s\ell} \le O\left(2C/t\right)^{s\ell}$. Then $\mathbb{E}q(\Sigma_z 1/2\boldsymbol{w}) \le (2Cs/t)^s$ and $\mathbb{E}q(\Sigma_z^{1/2}\boldsymbol{w})^2 \le (4Cs/t)^{2s}$. Therefore, by Chebyshev's inequality, with probability 0.99999,

$$q(\Sigma_z^{1/2}\boldsymbol{w}) \le (2Cs/t)^s + \sqrt{100000}\,(4Cs/t)^s \le 320\,(4Cs/t)^s\,.$$

In this case, we also have

$$q(\mu_a - \mu_b + \Sigma_z^{1/2}\boldsymbol{w}) \ge (c/16)^s - 320\,(4Cs/t)^s \ge (c/16)^s\,/2,$$

where in the last inequality we used that $t \ge 50000Cs/c$. This concludes the proof. ■

Lemma 17 (Existence of vector that satisfies moment contraints) *Let $s, t \ge \lceil \log p_{\min}^{-1} \rceil$ integers. If $t \le \max_{i,j} \|\Sigma_z^{-1/2}(\mu_i - \mu_j)\|^2$, there exists some $v \in \mathbb{R}^d$ that satisfies $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} = 1$ and $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \le 30^t$. Furthermore, $\|\operatorname{cov}(\boldsymbol{z})^{1/2} v\|^2 \le 8$.*

Proof Let $(a, b) = \arg\max_{(i,j)} \|\Sigma_z^{-1/2}(\mu_i - \mu_j)\|$ and let $v = \Sigma_z^{-1}(\mu_a - \mu_b)$. The vector for which we will guarantee the stated properties is $v^* = \frac{v}{(\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s})^{1/2s}}$.

We begin by proving that $\max_{i,j}\langle \mu_i - \mu_j, v\rangle^2 = \langle \mu_a - \mu_b, v\rangle^2$, which will be used later. We have

$$\begin{aligned}
\max_{i,j}\langle \mu_i - \mu_j, v\rangle^2 &= \max_{i,j}\langle \mu_i - \mu_j, \Sigma_z^{-1}(\mu_a - \mu_b)\rangle^2 \\
&= \max_{i,j}\langle \Sigma_z^{-1/2}(\mu_i - \mu_j), \Sigma_z^{-1/2}(\mu_a - \mu_b)\rangle^2 \\
&\le \|\Sigma_z^{-1/2}(\mu_i - \mu_j)\|^2 \cdot \|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^2 \\
&\le \|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^4 \\
&= \langle \mu_a - \mu_b, v\rangle^2.
\end{aligned}$$

We also have for the variance in direction v that

$$(v^\top \Sigma_z v)^t = ((\mu_a - \mu_b)^\top \Sigma_z^{-1}(\mu_a - \mu_b))^t = \|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{2t}.$$

We now derive upper bounds for the $2t$ moments of $\langle \boldsymbol{z}, v\rangle$ in direction v and lower bounds for the $2s$ moments in direction v. Recall that we assume $t \le \max_{i,j} \|\Sigma_z^{-1/2}(\mu_i - \mu_j)\|^2$. For the upper bound, using Lemma 13 we have

$$\begin{aligned}
\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} &\le 2^{2t-1}\mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2t} + 2^{2t-1}(v^\top \Sigma_z v)^t t^t \\
&\le 2^{2t-1}\max_{i,j}\langle \mu_i - \mu_j, v\rangle^{2t} + 2^{2t-1}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{2t} t^t \\
&= 2^{2t-1}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{4t} + 2^{2t-1}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{2t} t^t \\
&\le 2^{2t-1}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{4t} + 2^{2t-1}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{4t} \\
&\le 2^{2t}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{4t}.
\end{aligned}$$

For the lower bound, using Lemma 14 we have

$$\begin{aligned}
\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} &\geq \mathbb{E}\langle \boldsymbol{\mu}_z, v\rangle^{2s} + (v^\top \Sigma_z v)^s \frac{s^s}{2^s} \\
&\geq p_{\min}^2 \max_{i,j}\langle \mu_i - \mu_j, v\rangle^{2s} + \|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{2s} \frac{s^s}{2^s} \\
&= p_{\min}^2 \|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{4s} + \|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{2s} \frac{s^s}{2^s} \\
&\geq p_{\min}^2 \|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{4s}.
\end{aligned}$$

Recall that $v^* = \frac{v}{(\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s})^{1/2s}}$. Clearly, $\mathbb{E}\langle \boldsymbol{z}, v^*\rangle^{2s} = 1$. Furthermore,

$$\mathbb{E}\langle \boldsymbol{z}, v^*\rangle^{2t} = \frac{\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t}}{(\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s})^{t/s}} \leq \frac{2^{2t}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{4t}}{p_{\min}^{2t/s}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^{4t}} = \left(\frac{4}{p_{\min}^{2/s}}\right)^t \leq \left(4e^2\right)^t \leq 30^t$$

where in the last inequality we used that $p_{\min}^{1/s} \geq e^{-1}$. Therefore v^* satisfies the desired moment constraints.

Finally, we prove that $\|\operatorname{cov}(\boldsymbol{z})^{1/2}v^*\|^2 \leq 8$. Note that $\operatorname{cov}(\boldsymbol{z}) = \operatorname{cov}(\boldsymbol{\mu}_z) + \Sigma_z$. We have then

$$\begin{aligned}
\|\operatorname{cov}(\boldsymbol{z})^{1/2}v^*\|^2 &= (v^*)^\top \operatorname{cov}(\boldsymbol{z})v^* \\
&= (v^*)^\top \operatorname{cov}(\boldsymbol{\mu}_z)v^* + (v^*)^\top \Sigma_z v^* \\
&\overset{(1)}{\leq} \max_{i,j}\langle \mu_i - \mu_j, v^*\rangle^2 + \|\Sigma_z^{1/2}v^*\|^2 \\
&= \frac{\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^4}{(\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s})^{1/s}} + \frac{\|\Sigma_z^{1/2}\Sigma_z^{-1}(\mu_a - \mu_b)\|^2}{(\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s})^{1/s}} \\
&\overset{(2)}{\leq} \frac{\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^4}{p_{\min}^{2/s}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^4} + \frac{\|\Sigma_z^{1/2}\Sigma_z^{-1}(\mu_a - \mu_b)\|^2}{p_{\min}^{2/s}\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^4} \\
&\overset{(3)}{\leq} e^2 + e^2 \frac{1}{\|\Sigma_z^{-1/2}(\mu_a - \mu_b)\|^2} \\
&\overset{(4)}{=} e^2 + o(1) \leq 8,
\end{aligned}$$

where in (1) we used that $(v^*)^\top \operatorname{cov}(\boldsymbol{\mu}_z)v^* = (v^*)^\top \mathbb{E}\boldsymbol{\mu}_z\boldsymbol{\mu}_z^\top v^* \leq \max_{\mu_z}(v^*)^\top \mu_z\mu_z^\top v^*$ for μ_z in the support of $\boldsymbol{\mu}_z$, in (2) we used the lower bound that we derived above on $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s}$, in (3) we used that $p_{\min}^{1/s} \geq e^{-1}$, and in (4) we used mean separation.

■

B.2. Finite sample bounds

Recall that, to apply Lemma 12, we need to find a pseudo-expectation that satisfies the moment lower bound $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} \geq c^s$ and the moment upper bound $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq C^t$ for $s \ll t$. Lemma 18 shows that it suffices to find a pseudo-expectation that satisfies $\|\widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2}v\|^2 \lesssim 8$, $\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2s} \gtrsim c^s$, and $\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} \lesssim C^t$. Without the bound on the norm of $\widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2}v$ the errors may be arbitrarily large.

The result is supported by Lemma 19, which shows that quadratics in the empirical covariance matrix are close to quadratics in the population covariance matrix of the components, and by Lemma 20, which shows that the empirical moments are close to the population moments. The proofs of these two lemmas are deferred to the appendix.

Lemma 18 (Moment constraints from empirical moment constraints) *Let* $\eta < 0.001$. *Let* c, $C \geq 0$ *and let* $t \geq 1$ *integer. For* $n \geq (p_{\min}^{-1} d)^{O(t)} \eta^{-2} \epsilon^{-1}$, *with probability* $1 - \epsilon$,

$$\left\{ \|\widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2} v\|^2 \leq (1+\eta) \cdot 8, \hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2s} \geq c^s + \eta, \hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} \leq C^t - \eta \right\}$$
$$\Big|\frac{v}{2t} \left\{ \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} \geq c^s, \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq C^t \right\}.$$

Furthermore, with probability $1-\epsilon$, *the axiom is satisfied with* s, t, *and* v *as in Lemma 17 and with* $c = (1-\eta)^{1/s}$ *and* $C = (30^t + \eta)^{1/t}$.

Proof By Lemma 19, $\Big|\frac{v}{2} \|\operatorname{cov}(\boldsymbol{z})^{1/2} v\|^2 \leq (1+\eta)^2 \cdot 8 \leq 9$. Then, by Lemma 20, $\Big|\frac{v}{2s} \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} \geq c^s$ and $\Big|\frac{v}{2t} \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq C^t$. With n as given, this holds with probability at least $1-\epsilon$.

For the second claim of the lemma, we have by Lemma 17 that there exists some v with $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2s} = 1$, $\mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq 30^t$, and $\|\operatorname{cov}(\boldsymbol{z})^{1/2} v\|^2 \leq 8$. By Lemma 19, we also have that $\|\widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2} v\|^2 \leq (1+\eta)^2 \cdot 8 \leq 9$, and then by Lemma 20, we also have that $\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2s} \geq 1 - \eta$ and $\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} \leq 30^t + \eta$. Again, with n as given, this holds with probability at least $1-\epsilon$. ■

Lemma 19 *Let* $C \geq 0$. *For* $n \geq kd^2 \log^2(d/\epsilon) O(\eta^{-2})$, *with probability* $1-\epsilon$,

$$\{\|\widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2} v\|^2 \leq C\} \Big|\frac{v}{2} \{\|\operatorname{cov}(\boldsymbol{z})^{1/2} v\|^2 \leq (1+\eta)C\},$$

$$\{\|\operatorname{cov}(\boldsymbol{z})^{1/2} v\|^2 \leq C\} \Big|\frac{v}{2} \{\|\widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2} v\|^2 \leq (1+\eta)C\}.$$

Proof See Section F.5. ■

Lemma 20 *Let* $C \geq 0$ *and let* $t \geq 1$ *integer. For* $n \geq (Cp_{\min}^{-1} d)^{O(t)} \eta^{-2} \epsilon^{-1}$, *with probability* $1-\epsilon$,

$$\{\|\operatorname{cov}(\boldsymbol{z})^{1/2} v\|^2 \leq C\} \Big|\frac{v}{O(t)} \{\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} \leq \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} + \eta\},$$

$$\{\|\operatorname{cov}(\boldsymbol{z})^{1/2} v\|^2 \leq C\} \Big|\frac{v}{O(t)} \{\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} \geq \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} - \eta\}.$$

Proof See Section F.5. ■

B.3. Proof of Theorem 10

Proof Let $s = \lceil \log p_{\min}^{-1} \rceil$ and $t = 10000000s$. The algorithm is:

1. Compute a pseudo-expectation $\tilde{\mathbb{E}}$ of degree $2t$ over $v \in \mathbb{R}^d$ such that $\|\widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2} v\|^2 \leq 1.01 \cdot 8$, $\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2s} \geq 1 - 0.005$, and $\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} \leq 30^t + 0.005\}$.
2. Construct a separating polynomial q based on $\tilde{\mathbb{E}}$ as in Lemma 12.
3. Return q.

We now analyze the algorithm. First, we argue that there exists a pseudo-expectation $\tilde{\mathbb{E}}$ that satisfies the given constraints. Note that $t \leq \max_{i,j} \|\Sigma_z^{-1/2}(\mu_i - \mu_j)\|^2$ if C_{sep} is a large enough constant. Therefore, the conditions of Lemma 17 for s and t are satisfied. Then, by Lemma 18, for $n \geq n_0$, there exists a vector $v \in \mathbb{R}^d$ that satisfies the given constraints. Then, there also exists a pseudo-expectation that satisfies the constraints.

Second, we argue that q has the desired properties. By Lemma 18, $\tilde{\mathbb{E}}$ also sastisfies with high probability that $\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2s} \geq 1 - 0.01 \geq 0.99^s$ and $\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} \leq 30^t + 0.01 \leq 31^t$. Then the conditions of Lemma 12 are satisfied with $c = 0.99$ and $C = 31$. Then, we are guaranteed to return a separating polynomial q with the following properties:

- For independent random vectors $\boldsymbol{y}$ and $\boldsymbol{y}'$ sampled from different components, we have with probability at least 0.99999 that
$$q(\boldsymbol{y} - \boldsymbol{y}') \geq \frac{1}{2}\left(\frac{c}{16}\right)^s \geq \frac{1}{2}\left(\frac{0.99}{16}\right)^s \geq \frac{1}{20^s}.$$
- For independent random vectors $\boldsymbol{y}$ and $\boldsymbol{y}'$ sampled from the same component, we have with probability at least 0.99999 that
$$q(\boldsymbol{y} - \boldsymbol{y}') \leq 320\left(\frac{4Cs}{t}\right)^s \leq 320\left(\frac{4 \cdot 31 \cdot s}{t}\right)^s \leq \frac{1}{200^s}.$$

The time complexity is dominated by the time to compute the pseudo-expectation. The pseudo-expectation is of degree $O(\log p_{\min}^{-1})$ over d variables, and each constraint requires summing over the n samples. Therefore, the time to compute the pseudo-expectation is $n \cdot d^{O(\log p_{\min}^{-1})}$. ■

B.4. Proof of Theorem 11

We start by stating and proving Lemma 21, which shows how to obtain a bipartition of the samples given a suitable distance function. We then prove Theorem 11.

Lemma 21 (Bipartition from distance function) *Assume access to a distance function $d_q : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that:*

- *There exist distinct $a, b \in [k]$ such that the independent random vectors $\boldsymbol{y} \sim N(\mu_a, \Sigma)$ and $\boldsymbol{y}' \sim N(\mu_b, \Sigma)$ satisfy*
$$\mathbb{P}\left\{d_q(\boldsymbol{y}, \boldsymbol{y}') \geq \frac{1}{\sqrt{20}}\right\} \geq 0.99999.$$

- *For all $a \in [k]$, the independent random vectors $\boldsymbol{y}, \boldsymbol{y}' \sim N(\mu_a, \Sigma)$ satisfy*

$$\mathbb{P}\left\{d_q(\boldsymbol{y}, \boldsymbol{y}') \leq \frac{1}{\sqrt{200}}\right\} \geq 0.99999.$$

Given a sample of size $n = \Omega(p_{\min}^{-1})$ from the mixture, there exists a polynomial-time algorithm that returns with probability 0.99 a partition of $[n]$ into two sets C_1 and C_2 such that, if true clustering of the samples is $S_1, ..., S_k$, then

$$\max_i \frac{|C_1 \cap S_i|}{|S_i|} \geq 0.99 \quad \textit{and} \quad \max_i \frac{|C_2 \cap S_i|}{|S_i|} \geq 0.99.$$

Proof The algorithm is:

1. Choose $i \in [n]$ uniformly at random.
2. Let $S = \{j \in [n] : d_q(y_i, y_j) \leq \frac{1}{\sqrt{200}}\}$.
3. Return S and $[n] \setminus S$.

We now analyze the algorithm.

First, we prove that, with probability at least 0.995, a 0.99-fraction of the samples from the same component as i are included in S. With high probability, a 0.99998-fraction of the pairs of samples (y, y') with y and y' from the same component as i satisfy $d_q(y, y') \leq \frac{1}{\sqrt{200}}$. Then, the fraction of samples from this component that are farther than $\frac{1}{\sqrt{200}}$ from more than a 0.01-fraction of the other samples in the component is at most $\frac{1-0.99998}{0.01} = 0.002$. Then, overall, with probability at least 0.995, i is closer than $\frac{1}{\sqrt{200}}$ to at least a 0.99-fraction of the other samples in the component. In this case, S includes a 0.99-fraction of the samples from the same component as i.

Second, we prove that, with probability at least 0.995, at least a 0.99-fraction of the samples from one of the components are not included in S. Let $a, b \in [k]$ be the two components for which the large-distance guarantee holds. With high probability, a 0.99998-fraction of the pairs of samples (y, y') with y from a and y' from b satisfy $d_q(y, y') \geq \frac{1}{\sqrt{20}}$. We have $d_q(y, y') \leq d_q(y, y_i) + d_q(y', y_i)$, so if $d_q(y, y') \geq \frac{1}{\sqrt{20}}$, then at least one of $d_q(y, y_i)$ or $d_q(y', y_i)$ is at least $\frac{1}{2\sqrt{20}} > \frac{1}{\sqrt{200}}$. Then, for such pairs, it is impossible for both y and y' to be in S. Suppose that a p_a-fraction of the samples from a are in S and that a p_b-fraction of the samples from b are in S. We need then that $1 - p_a p_b \geq 0.99998$, so $\min(p_a, p_b) \leq \sqrt{1 - 0.9998} \leq 0.01$. Therefore, $[n] \setminus S$ contains at least a 0.99-fraction of the samples from one of a or b.

Therefore, with probability at least 0.99, the conclusion of the lemma holds. ■

Proof of Theorem 11 The algorithm is:

1. Run the algorithm from Theorem 10 to obtain a polynomial q.
2. Run the algorithm from Lemma 21 with the distance function $d_q(x, y) = q(x - y)^{1/2s}$.
3. Return the resulting bipartition.

We now analyze the algorithm. Recall that $q(u) = \tilde{\mathbb{E}}\langle u, v\rangle^{2s}$, so $q(u)^{1/2s} = (\tilde{\mathbb{E}}\langle u, v\rangle^{2s})^{1/2s}$. Then, we have byLemma 62 that $q^{1/2s}$ satisfies the triangle inequality. It follows that $d_q(x, y) = q(x - y)^{1/2s}$ is a distance function. Then, by the guarantees of Theorem 10, d_q satisfies the requirements of Lemma 21, so the stated guarantees follow.

The time complexity is dominated by the time complexity of the algorithm from Theorem 10.

■

Appendix C. Parallel pancakes

The model studied in this section is a well-separated mixture of Gaussians with colinear means that is in isotropic position. As shown in Section C.1, the isotropic position property makes this model similar to the parallel pancakes construction, in the sense that the only direction in which the components of the mixture have variance different from 1 is the direction of the means. In Section D we study the same model without the isotropic position assumption.

Setting. We consider a mixture of k Gaussian distributions $N(\mu_i, \Sigma)$ with mixing weights p_i for $i = 1, ..., k$, where $\mu_i \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite, and $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$. Let $p_{\min} = \min_i p_i$.

The distribution is in isotropic position: for $\boldsymbol{y}$ distributed according to the mixture, we have $\mathbb{E}\boldsymbol{y} = 0$ and $\operatorname{cov}(\boldsymbol{y}) = I_d$.

The distribution also satisfies mean separation and mean colinearity:

- Mean separation: for some $C_{sep} > 0$ and for all $i \neq j$,

$$\left\|\Sigma^{-1/2}(\mu_i - \mu_j)\right\|^2 \geq C_{sep} \log p_{\min}^{-1}.$$

- Mean colinearity: for some unit vector $u \in \mathbb{R}^d$ and for all i,

$$\mu_i = \langle \mu_i, u\rangle u.$$

Also define $\sigma^2 = u^\top \Sigma u$, which is the variance of the components in the direction of the means.

Theorem 22 (Parallel pancakes algorithm) *Consider the Gaussian mixture model defined above, with C_{sep} larger than some universal constant. Let*

$$n_0 = \left(\frac{1}{\sigma^2}\right)^{O(1)} \cdot (p_{\min}^{-1} d)^{O(\log p_{\min}^{-1})}.$$

Given a sample of size $n \geq n_0$ from the mixture, there exists an algorithm that runs in time $n^{O(\log p_{\min}^{-1})}$ and returns with high probability a partition of $[n]$ into k sets $C_1, ..., C_k$ such that, if the true clustering of the samples is $S_1, ..., S_k$, then there exists a permutation π of $[k]$ such that

$$1 - \frac{1}{n}\sum_{i=1}^{k} |C_i \cap S_{\pi(i)}| \leq \left(\frac{p_{\min}}{k}\right)^{O(1)}.$$

We introduce some further notation for this section. Let $\boldsymbol{y}$ be distributed according to the mixture. We specify the model as $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{w}$, where $\boldsymbol{\mu}$ takes value μ_i with probability p_i and $\boldsymbol{w} \sim N(0, \Sigma)$, with $\boldsymbol{\mu}$ and $\boldsymbol{w}$ independent of each other.

C.1. Isotropic position properties

In this section we prove some consequences of the fact that $\boldsymbol{y}$ is in isotropic position. Lemma 23 shows that $\Sigma = I_d - (1-\sigma^2)uu^\top$ with $0 < \sigma^2 \leq 1$. This means that Σ can have at most one eigenvalue less than 1 and that the eigenvectors corresponding to this eigenvalue are parallel to the direction of the means u. Then Lemma 24 uses this form of Σ to quantify the separation of the means along direction u in terms of σ^2.

Lemma 23 (Isotropic position component covariance matrix) *We have that:*

1. $\sigma^2 = 1 - \sum_{i=1}^k p_i \langle \mu_i, u\rangle^2$,
2. $\Sigma = I_d - (1-\sigma^2)uu^\top$, *and (3)* $0 < \sigma^2 \leq 1$.

Proof We have that $\mathbb{E}\boldsymbol{y} = \mathbb{E}\boldsymbol{\mu} + \mathbb{E}\boldsymbol{w} = \mathbb{E}\boldsymbol{\mu}$. Because the distribution is in isotropoic position, we also have that $\mathbb{E}\boldsymbol{y} = 0$, so the equation above implies that $\mathbb{E}\boldsymbol{\mu} = 0$. Then $\operatorname{cov}(\boldsymbol{\mu}) = \sum_{i=1}^k p_i \mu_i \mu_i^\top$.

Furthermore, since $\boldsymbol{\mu}$ and $\boldsymbol{w}$ are independent, we have that $\operatorname{cov}(\boldsymbol{y}) = \operatorname{cov}(\boldsymbol{\mu}) + \operatorname{cov}(\boldsymbol{w}) = \sum_{i=1}^k p_i \mu_i \mu_i^\top + \Sigma$. Because the distribution is in isotropic position, we also have that $\operatorname{cov}(\boldsymbol{y}) = I_d$, so the equation above implies $\Sigma = I_d - \sum_{i=1}^k p_i \mu_i \mu_i^\top$. Plugging in $\mu_i = \langle \mu_i, u\rangle u$, we have $\Sigma = I_d - \left(\sum_{i=1}^k p_i \langle \mu_i, u\rangle^2\right) uu^\top$.

Then, it follows that $\sigma^2 = u^\top \Sigma u = 1 - \sum_{i=1}^k p_i \langle \mu_i, u\rangle^2$. This proves (1). The fact that $1 - \sigma^2 = \sum_{i=1}^k p_i \langle \mu_i, u\rangle^2$ also proves (2). For (3), $\sigma^2 > 0$ follows by the definition using that Σ is positive definite and $\sigma^2 \leq 1$ follows by (1). ∎

Lemma 24 (Isotropic position mean separation) *For all $i \neq j$,*

$$\langle \mu_i - \mu_j, u\rangle^2 \geq C_{sep} \sigma^2 \log p_{\min}^{-1}.$$

Proof By Lemma 23, $\Sigma = I_d - (1-\sigma^2)uu^\top$. This implies that $\Sigma^{-1/2} = I_d + \left(1/\sqrt{\sigma^2} - 1\right) uu^\top$. Then, using that $\mu_i = \langle \mu_i, u\rangle u$, we have that

$$\Sigma^{-1/2}(\mu_i - \mu_j) = \left(I_d + \left(\frac{1}{\sqrt{\sigma^2}} - 1\right) uu^\top\right)(\langle \mu_i - \mu_j, u\rangle u) = \frac{1}{\sqrt{\sigma^2}} \langle \mu_i - \mu_j, u\rangle u.$$

Therefore, the separation condition $\left\|\Sigma^{-1/2}(\mu_i - \mu_j)\right\|^2 \geq C_{sep} \log p_{\min}^{-1}$ is equivalent to $\frac{1}{\sigma^2}\langle \mu_i - \mu_j, u\rangle^2 \geq C_{sep} \log p_{\min}^{-1}$. The conclusion follows by multiplying both sides by σ^2. ∎

C.2. Exact moment direction recovery

In this section we discuss how to recover a direction close to the direction of the means u, assuming oracle access to moments $\mathbb{E}\boldsymbol{y}^{\otimes t}$ for any positive integer t. Access to these moments allows us to calculate exactly directional moments of the form $\mathbb{E}\langle \boldsymbol{y}, v\rangle^t$, which simplifies the analysis. Finite sample considerations are discussed starting with Section C.3.

Theorem 25 shows that there exists an algorithm that computes a unit vector $\hat{u}$ with correlation $1 - O(\min(\sigma^2, \frac{1}{k}))$ with the direction of the means u. We remark that it is necessary for $\hat{u}$ to have a correlation of at least $1 - O(\sigma^2)$ with u in order for the components of the mixture to be separated along direction $\hat{u}$.

Theorem 25 (Direction recovery with exact moments) *Assume oracle access to $\mathbb{E}\boldsymbol{y}^{\otimes t}$ for any positive integer t. Then there exists an algorithm with time complexity $\left(\log \frac{1}{\sigma^2}\right) \cdot d^{O(\log p_{\min}^{-1})}$ that outputs a unit vector $\hat{u} \in \mathbb{R}^d$ such that $\langle u, \hat{u}\rangle^2 \geq 1 - 320 \min(\sigma^2, \frac{1}{k})$.*

The two main ingredients for Theorem 25 are Theorem 26, which gives an algorithm to compute pseudo-expectations over unit vectors correlated with u, and Theorem 27, which gives an algorithm to sample from such pseudo-expectations. We note that Theorem 26 can be interpreted as a collection of sum-of-squares identifiability proofs for the direction of the means u.

We state these two supporting theorems and then prove Theorem 25. After that, we work toward proving the supporting theorems.

Theorem 26 (Direction sum-of-squares identifiability) *Assume oracle access to $\mathbb{E}\boldsymbol{y}^{\otimes t}$ for any positive integer t. Then there exists an algorithm with time complexity $\left(\log \frac{1}{\sigma^2}\right) \cdot d^{O(\log p_{\min}^{-1})}$ that computes two peseudo-expectations $\tilde{\mathbb{E}}_U$ and $\tilde{\mathbb{E}}_L$ of degree $O(\log p_{\min}^{-1})$ over a variable $v \in \mathbb{R}^d$ such that the following holds. Let $s = \lceil \log p_{\min}^{-1} \rceil$, let $t = 5000s$, and let $\tau = \frac{800e}{C_{sep}k^2}$. Then $\tilde{\mathbb{E}}_U \|v\|^2 = 1$, $\tilde{\mathbb{E}}_L \|v\|^2 = 1$, and:*

- *If $\sigma^2 \geq \tau$, then $\tilde{\mathbb{E}}_U \langle u, v\rangle^{2s} \geq (1-\tau)^s$.*
- *If $\sigma^2 < \tau$ and $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \geq (4es)^s$, then $\tilde{\mathbb{E}}_U \langle u, v\rangle^{2s} \geq (1-\sigma^2)^s$.*
- *If $\sigma^2 < 0.001$ and $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \leq (100s)^s$, then $\tilde{\mathbb{E}}_L \langle u, v\rangle^{2t} \geq (1-20\sigma^2)^t$.*

Theorem 27 (Direction sum-of-squares sampling) *Let $t \in \mathbb{N}$ and $\epsilon \in \mathbb{R}$ such that $t \geq 1$ and $0 \leq \epsilon \leq 1/(3t^2)$. Let $u \in \mathbb{R}^d$ be a unit vector. Given a pseudo-expectation $\tilde{\mathbb{E}}$ of degree $2t$ over a variable $v \in \mathbb{R}^d$ that satisfies $\tilde{\mathbb{E}}\|v\|^2 = 1$ and $\tilde{\mathbb{E}}\langle u, v\rangle^{2t} \geq (1-\epsilon)^t$, there exists an algorithm with time complexity $d^{O(t)}$ that returns a unit vector $\hat{u} \in \mathbb{R}^d$ such that $\langle u, \hat{u}\rangle^2 \geq 1 - 16\epsilon$.*

Proof of Theorem 25 Let $\tau = \frac{800e}{C_{sep}k^2}$. The algorithm is:

1. Run the algorithm from Theorem 26 to obtain pseudo-expectations $\tilde{\mathbb{E}}_U$ and $\tilde{\mathbb{E}}_L$.
2. Run the algorithm from Theorem 27 for pseudo-expectations $\tilde{\mathbb{E}}_U$ and $\tilde{\mathbb{E}}_L$ to obtain unit vectors $\hat{u}_U \in \mathbb{R}^d$ and $\hat{u}_L \in \mathbb{R}^d$, respectively.
3. If $\sigma^2 \geq \tau$, return $\hat{u}_U$. Else, for $s = \lceil \log p_{\min}^{-1} \rceil$, if $\mathbb{E}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s} \geq (50s)^s$, return $\hat{u}_U$. Else, return $\hat{u}_L$.

We now analyze the algorithm. We consider the three possible cases in step (3) of the algorithm:

- Suppose $\sigma^2 \geq \tau$. Then Theorem 26 guarantees that $\tilde{\mathbb{E}}_U \langle u, v\rangle^{2s} \geq (1-\tau)^s$, so by Theorem 27 we have $\langle u, \hat{u}_U\rangle^2 \geq 1 - 16\tau \geq 1 - 16\min(\sigma^2, \tau)$.

- Suppose $\sigma^2 < \tau$ and $(\mathbb{E}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s})^{1/s} \geq 50s$. By Lemma 29, $(\mathbb{E}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s})^{1/s} \leq (\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s} + es$, so it must be the case that $(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s} \geq (50-e)s \geq 4es$. Then Theorem 26 guarantees that $\tilde{\mathbb{E}}_U \langle u, v\rangle^{2s} \geq (1-\sigma^2)^s$, so by Theorem 27 we have $\langle u, \hat{u}_U\rangle^2 \geq 1 - 16\sigma^2 \geq 1 - 16\min(\sigma^2, \tau)$.

- Suppose $\sigma^2 < \tau$ and $\mathbb{E}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s} < (50s)^s$. We have by Lemma 30 that $(\mathbb{E}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s})^{1/s} \geq (\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s}$, so it must be the case that $(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s} < 50s \leq 100s$. Then Theorem 26 guarantees that $\tilde{\mathbb{E}}_L \langle u, v\rangle^{2t} \geq (1-20\sigma^2)^t$, so by Theorem 27 we have $\langle u, \hat{u}_L\rangle^2 \geq 1 - 16 \cdot 20\sigma^2 \geq 1 - 320\min(\sigma^2, \tau)$.

Let $\hat{u}$ be the unit vector returned by step (3) of the algorithm. Then we are guaranteed that in all cases $\langle u, \hat{u}\rangle^2 \geq 1 - 320\min(\sigma^2, \tau) \geq 1 - 320\min(\sigma^2, \frac{1}{k})$, where we used the loose upper bound $\tau \leq \frac{1}{k}$.

The time complexity of the algorithm is dominated by the time to run the algorithm from Theorem 26. ■

C.2.1. SUM-OF-SQUARES IDENTIFIABILITY (PROOF OF THEOREM 26)

We prove a number of supporting lemmas and then prove Theorem 26. The most important components are Lemma 32 and Lemma 33, which give sum-of-squares proofs that, for suitably chosen $s, t = O(\log p_{\min}^{-1})$, either the maximizer of $\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2s}$ or the minimizer of $\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}$ over unit vectors v must be close to u.

We start with Lemma 28, Lemma 29 and Lemma 30, which give sum-of-squares bounds on the moments of the mixture. Informally, for $t = \Omega(\log p_{\min}^{-1})$, these bounds correspond to the following decomposition of the directional $2t$ moments:

$$(\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t})^{1/t} = \langle u, v\rangle^2 \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + \Theta(1) \cdot t(v^\top \Sigma v). \tag{18}$$

Lemma 28 (Moment equality) *For $t \geq 1$ integer,*

$$\vdash_{2t}^{v} \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} = \sum_{s=0}^{t} \binom{2t}{2s} \langle u, v\rangle^{2s} \mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} (v^\top \Sigma v)^{t-s} (2t-2s-1)!!.$$

Proof

$$\begin{aligned}
\vdash_{2t}^{v} \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} &= \mathbb{E}(\langle \boldsymbol{\mu}, v\rangle + \langle \boldsymbol{w}, v\rangle)^{2t} = \sum_{j=0}^{2t} \binom{2t}{j} \mathbb{E}\langle \boldsymbol{\mu}, v\rangle^{j} \langle \boldsymbol{w}, v\rangle^{2t-j} \\
&\overset{(1)}{=} \sum_{j=0}^{2t} \binom{2t}{j} \mathbb{E}\langle \boldsymbol{\mu}, v\rangle^{j} \mathbb{E}\langle \boldsymbol{w}, v\rangle^{2t-j} \overset{(2)}{=} \sum_{s=0}^{t} \binom{2t}{2s} \mathbb{E}\langle \boldsymbol{\mu}, v\rangle^{2s} \mathbb{E}\langle \boldsymbol{w}, v\rangle^{2t-2s} \\
&\overset{(3)}{=} \sum_{s=0}^{t} \binom{2t}{2s} \langle u, v\rangle^{2s} \mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} (v^\top \Sigma v)^{t-s} (2t-2s-1)!!
\end{aligned}$$

where in (1) we used that $\boldsymbol{\mu}$ and $\boldsymbol{w}$ are independent, in (2) we used that $\mathbb{E}\langle \boldsymbol{w}, v\rangle^{2t-j} = 0$ for $2t - j$ odd, and in (3) we used that $\boldsymbol{\mu} = \langle \boldsymbol{\mu}, u\rangle u$ and that $\mathbb{E}\langle \boldsymbol{w}, v\rangle^{2t-2s} = (v^\top \Sigma v)^{t-s}(2t-2s-1)!!$. ■

Lemma 29 (Moment upper bound) *For $t \geq 1$ integer,*

$$\vdash_{2t}^{v} \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} \leq \left(\langle u, v\rangle^2 \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + et(v^\top \Sigma v)\right)^t.$$

Proof Starting with the result in Lemma 28,

$$\begin{aligned}
\left|\frac{v}{2t}\right. \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} &= \sum_{s=0}^{t} \binom{2t}{2s} \langle u, v\rangle^{2s} \mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} (v^\top \Sigma v)^{t-s} (2t-2s-1)!! \\
&\overset{(1)}{\leq} \sum_{s=0}^{t} \binom{2t}{2s} \langle u, v\rangle^{2s} \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{s/t} (v^\top \Sigma v)^{t-s} (2t-2s-1)!! \\
&\overset{(2)}{\leq} \sum_{s=0}^{t} \binom{t}{s} \langle u, v\rangle^{2s} \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{s/t} (v^\top \Sigma v)^{t-s} (et)^{t-s} \\
&= \left(\langle u, v\rangle^{2} \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + et(v^\top \Sigma v)\right)^{t}.
\end{aligned}$$

In (1) we used that $s \leq t$ and Jensen's inequality as follows:

$$\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} = \mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t(s/t)} \leq \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{s/t}.$$

In (2) we used that $\binom{2t}{2s}(2t-2s-1)!! \leq \binom{t}{s}(et)^{t-s}$ for $0 \leq s \leq t$ integers, which is proved in Lemma 75. ■

Lemma 30 (Moment lower bound) *For $t \geq 1$ integer,*

$$\left|\frac{v}{2t}\right. \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} \geq \left(\langle u, v\rangle^{2} \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + p_{\min}^{1/t} t/2 (v^\top \Sigma v)\right)^{t}.$$

Proof Starting with the result in Lemma 28,

$$\begin{aligned}
\left|\frac{v}{2t}\right. \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} &= \sum_{s=0}^{t} \binom{2t}{2s} \langle u, v\rangle^{2s} \mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} (v^\top \Sigma v)^{t-s} (2t-2s-1)!! \\
&\overset{(1)}{\geq} \sum_{s=0}^{t} \binom{2t}{2s} \langle u, v\rangle^{2s} \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{s/t} \left(p_{\min}^{1/t}\right)^{t-s} (v^\top \Sigma v)^{t-s} (2t-2s-1)!! \\
&\overset{(2)}{\geq} \sum_{s=0}^{t} \binom{t}{s} \langle u, v\rangle^{2s} \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{s/t} \left(p_{\min}^{1/t}\right)^{t-s} (v^\top \Sigma v)^{t-s} (t/2)^{t-s} \\
&= \left(\langle u, v\rangle^{2} \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + p_{\min}^{1/t} t/2 (v^\top \Sigma v)\right)^{t}.
\end{aligned}$$

In (1) we used that $s \leq t$ and the fact that the s-norm is greater than or equal to the t-norm as follows:

$$\begin{aligned}
\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} &= \sum_{i=1}^{k} p_i \langle \mu_i, u\rangle^{2s} = \sum_{i=1}^{k} (p_i^{1/s} \langle \mu_i, u\rangle^{2})^{s} \\
&\geq \left(\sum_{i=1}^{k} (p_i^{1/s} \langle \mu_i, u\rangle^{2})^{t}\right)^{s/t} = \left(\sum_{i=1}^{k} p_i^{t/s} \langle \mu_i, u\rangle^{2t}\right)^{s/t} \\
&\geq \left(p_{\min}^{t/s-1} \sum_{i=1}^{k} p_i \langle \mu_i, u\rangle^{2t}\right)^{s/t} = p_{\min}^{1-s/t} \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{s/t}.
\end{aligned}$$

In (2) we used that $\binom{2t}{2s}(2t-2s-1)!! \geq \binom{t}{s}(t/2)^{t-s}$ for $0 \leq s \leq t$ integers, which is proved in Lemma 75.

■

Lemma 31 shows that the contribution of the means to the $\Omega(\log p_{\min}^{-1})$ moments in direction u is lower bounded by $\Omega(k^2\sigma^2 \log p_{\min}^{-1})$. This result is used in some of the later proofs to argue that if the mean contribution is small, then σ^2 is small, and conversely, that if σ^2 is large, then the mean contribution is large.

Lemma 31 *For $2s \geq \lceil \log p_{\min}^{-1} \rceil$ integer,*

$$(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s} \geq \frac{C_{sep}}{100} k^2 \sigma^2 \log p_{\min}^{-1}.$$

Proof By Lemma 24, for all $i \neq j$, $|\langle \mu_i - \mu_j, u\rangle| \geq \sqrt{C_{sep}\sigma^2 \log p_{\min}^{-1}}$. Then there exist $a, b \in [k]$ such that $|\langle \mu_a - \mu_b, u\rangle| \geq (k-1)\sqrt{C_{sep}\sigma^2 \log p_{\min}^{-1}}$. Hence, there exists $a \in [k]$ such that $|\langle \mu_a, u\rangle| \geq \frac{k-1}{2}\sqrt{C_{sep}\sigma^2 \log p_{\min}^{-1}}$. Then

$$(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s} \geq p_{\min}^{1/s} \max_i \langle \mu_i, u\rangle^2 \geq p_{\min}^{1/s} \left(\frac{k-1}{2}\right)^2 C_{sep}\sigma^2 \log p_{\min}^{-1} \geq \frac{C_{sep}}{100} k^2 \sigma^2 \log p_{\min}^{-1},$$

where we used that $p_{\min}^{1/s} \geq e^{-2}$. ■

We now state and prove the sum-of-squares identifiability proofs of Lemma 32 and Lemma 33. Let $s, t = O(\log p_{\min}^{-1})$ with $s \ll t$. Lemma 32 proves that, in the case $(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s} \geq \Theta(s)$, if $\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2s}$ is close to its maximum value over unit vectors v, then $\langle u, v\rangle^{2s}$ is close to 1. Lemma 33 proves that, in the opposite case $(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s} \leq \Theta(s)$, if $\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}$ is close to its minimum value over unit vectors v, then $\langle u, v\rangle^{2t}$ is close to 1.

Lemma 32 (Direction sum-of-squares identifiability from moment maximization) *Let $M \geq 2$. Let s be an integer such that $2s \geq \lceil \log p_{\min}^{-1} \rceil$. Suppose that $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \geq (Mes)^s$. Then, for $\epsilon \leq \sigma^2/M$,*

$$\left\{ \|v\|^2 = 1, \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2s} \geq (1-\epsilon)\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} - \epsilon\right) \right\} \vdash_{2s}^{v} \left\{ \langle u, v\rangle^{2s} \geq (1 - 4\sigma^2/M)^s \right\}.$$

Furthermore, $v = u$ satisfies the axiom with $\epsilon = 0$.

Proof Substitute the upper bound of Lemma 3 into the axiom:

$$\vdash_{2s}^{v} \left(\langle u, v\rangle^2 \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s}\right)^{1/s} + es(v^\top \Sigma v) \right)^s \geq (1-\epsilon)\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} - \epsilon\right).$$

Divide by $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s}$:

$$\vdash_{2s}^{v} \left(\langle u, v\rangle^2 + (v^\top \Sigma v)\frac{es}{(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s}} \right)^s \geq (1-\epsilon)\left(1 - \frac{\epsilon}{\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s}}\right).$$

Recall that $\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s}\right)^{1/s} \geq Mes$, and substitute the lower bound on both sides:

$$\vdash^{v}_{2s} \left(\langle u, v\rangle^2 + \frac{1}{M}(v^\top \Sigma v)\right)^s \geq (1-\epsilon)\left(1 - \frac{\epsilon}{(Mes)^s}\right)$$

Use that $v^\top \Sigma v = 1 - (1-\sigma^2)\langle u, v\rangle^2$:

$$\vdash^{v}_{2t} \left(\langle u, v\rangle^2 + \frac{1}{M}(1 - (1-\sigma^2)\langle u, v\rangle^2)\right)^s \geq (1-\epsilon)\left(1 - \frac{\epsilon}{(Mes)^s}\right).$$

We simplify now the right-hand side. Use the loose bound $1 - \epsilon/(Mes)^s \geq 1 - \epsilon \geq (1-\epsilon)^{s-1}$ to obtain

$$\vdash^{v}_{2s} \left(\langle u, v\rangle^2 + \frac{1}{M}(1 - (1-\sigma^2)\langle u, v\rangle^2)\right)^s \geq (1-\epsilon)^s.$$

Finally, apply Lemma 65 with $x = \langle u, v\rangle$ and $\gamma = \frac{1}{1-\epsilon}$ to obtain that

$$\begin{aligned}\vdash^{v}_{2s} \langle u, v\rangle^{2s} &\geq \left(\frac{M - \frac{1}{1-\epsilon}}{\frac{1}{1-\epsilon}} \frac{1}{M - 1 + \sigma^2}\right)^s = \left(\frac{M - 1 - M\epsilon}{M - 1 + \sigma^2}\right)^s \\ &\geq \left(\frac{M - 1 - \sigma^2}{M - 1 + \sigma^2}\right)^s \geq \left(1 - 4\sigma^2/M\right)^s.\end{aligned}$$

To show that $v = u$ satisfies the axiom, simply note that Lemma 33 implies that $\mathbb{E}\langle \boldsymbol{y}, u\rangle^{2s} \geq \mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s}$.

■

Lemma 33 (Direction sum-of-squares identifiability from moment minimization) *Suppose* $\sigma^2 < 0.001$. *Let* s *be an integer such that* $2s \geq \lceil \log p_{\text{min}}^{-1} \rceil$. *Suppose that* $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \leq (100s)^s$. *Let* t *be an integer such that* $t \geq 5000s$. *Then, for* $\epsilon \leq \sigma^2/100$,

$$\left\{\|v\|^2 = 1, \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} \leq (1+\epsilon)\left(\left(\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + et\sigma^2\right)^t + \epsilon\right)\right\} \vdash^{v}_{2t} \left\{\langle u, v\rangle^{2t} \geq (1 - 20\sigma^2)^t\right\}.$$

Furthemore, $v = u$ *satisfies the axiom with* $\epsilon = 0$.

Proof We start by proving that, for $t \geq s$, $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t} \leq (100e^2 s)^t$. We have that

$$p_{\text{min}} \cdot \max_i \langle \mu_i, u\rangle^{2s} \leq \mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \leq (100s)^s.$$

Taking the s-th root and using that $p_{\text{min}}^{-1/s} \leq e^2$, we obtain that $\max_i \langle \mu_i, u\rangle^2 \leq 100e^2 s$. Therefore, $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t} \leq \max_i \langle \mu_i, u\rangle^{2t} = (100e^2 s)^t$.

We now proceed with the main claim. Substitute the lower bound of Lemma 4 into the axiom:

$$\vdash^{v}_{2t} \left(\langle u, v\rangle^2 \left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + p_{\text{min}}^{1/t} t/2 (v^\top \Sigma v)\right)^t \leq (1+\epsilon)\left(\left(\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + \sigma^2 et\right)^t + \epsilon\right).$$

Divide by $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}$:

$$\Big|\!\frac{v}{2t} \left(\langle u, v\rangle^2 + (v^\top \Sigma v)\frac{p_{\min}^{1/t} t/2}{(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t})^{1/t}}\right)^t \leq (1+\epsilon)\left(\left(1+\sigma^2\frac{et}{(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t})^{1/t}}\right)^t + \frac{\epsilon}{\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}}\right).$$

Let $\Delta = \frac{p_{\min}^{1/t} t/2}{(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t})^{1/t}}$. Then

$$\Big|\!\frac{v}{2t} \left(\langle u, v\rangle^2 + \Delta(v^\top \Sigma v)\right)^t \leq (1+\epsilon)\left(\left(1+\sigma^2\frac{2e}{p_{\min}^{1/t}}\Delta\right)^t + \frac{\epsilon}{\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}}\right).$$

Note that $\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} \leq 100e^2 s$ and $p_{\min}^{1/t} = e^{-s/t} \geq e^{-1}$. Then $\Delta \geq \frac{e^{-1}t/2}{100e^2 s}$. For $t \geq 5000s$ we have then $\Delta \geq 10$ and $p_{\min}^{-1/t} \leq 1.4$. Then $\frac{2e}{p_{\min}^{1/t}}\Delta \leq 8\Delta$. Then:

$$\Big|\!\frac{v}{2t} \left(\langle u, v\rangle^2 + \Delta(v^\top \Sigma v)\right)^t \leq (1+\epsilon)\left(\left(1+8\Delta\sigma^2\right)^t + \frac{\epsilon}{\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}}\right).$$

Divide by $\left(1+8\Delta\sigma^2\right)^t$:

$$\Big|\!\frac{v}{2t} \left(\frac{\langle u, v\rangle^2 + \Delta(v^\top \Sigma v)}{1+8\Delta\sigma^2}\right)^t \leq (1+\epsilon)\left(1 + \frac{\epsilon}{\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\left(1+8\Delta\sigma^2\right)^t}\right).$$

Use that $v^\top \Sigma v = 1-(1-\sigma^2)\langle u, v\rangle^2$:

$$\Big|\!\frac{v}{2t} \left(\frac{\langle u, v\rangle^2 + \Delta(1-(1-\sigma^2)\langle u, v\rangle^2)}{1+8\Delta\sigma^2}\right)^t \leq (1+\epsilon)\left(1 + \frac{\epsilon}{\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\left(1+8\Delta\sigma^2\right)^t}\right).$$

We simplify now the term involving ϵ. Note that, by Jensen's inequality, $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t} \geq (\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^2)^t = (1-\sigma^2)^t$. Also note that $(1-\sigma^2)(1+8\Delta\sigma^2) \geq 1$ for $\sigma^2 \leq 1/2$ and $\Delta \geq 10$. Then use the loose bound

$$1 + \frac{\epsilon}{\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\left(1+8\Delta\sigma^2\right)^t} \leq 1 + \frac{\epsilon}{((1-\sigma^2)(1+8\Delta\sigma^2))^t} \leq 1+\epsilon \leq (1+\epsilon)^{t-1}$$

to obtain

$$\Big|\!\frac{v}{2t} \left(\frac{\langle u, v\rangle^2 + \Delta(1-(1-\sigma^2)\langle u, v\rangle^2)}{1+8\Delta\sigma^2}\right)^t \leq (1+\epsilon)^t.$$

Finally, apply Lemma 66 with $x = \langle u, v\rangle$ and $\gamma = \frac{1}{1+\epsilon}$ to obtain that

$$\begin{aligned}\Big|\!\frac{v}{2t}\ \langle u, v\rangle^{2t} &\geq \left(\frac{\frac{1}{1+\epsilon}\Delta - 1}{\frac{1}{1+\epsilon}(\Delta-1)}(1-10\sigma^2)\right)^t = \left(\left(1-\frac{\epsilon}{\Delta-1}\right)(1-10\sigma^2)\right)^t \\ &= \left((1-\sigma^2/100)(1-10\sigma^2)\right)^t \geq \left(1-20\sigma^2\right)^t.\end{aligned}$$

To show that $v = u$ satisfies the axiom, simply note that Lemma 32 implies that $\mathbb{E}\langle \boldsymbol{y}, u\rangle^{2t} \leq \left(\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + et\sigma^2\right)^t$.

■

We now prove Theorem 26.

Proof of Theorem 26 Let $s = \lceil \log p_{\min}^{-1} \rceil$, $t = 5000s$, and $\tau = 800e/(C_{sep}k^2)$. The algorithm is:

1. If $\sigma^2 \geq \tau$, then let $M = C_{sep}k^2\sigma^2/(200e)$. Else, let $M = 4$.
2. Binary search up to resolution $\sigma^2/(100M)$ the largest T_U in the interval $[0, (p_{\min}^{-1})^s]$ such that there exists a degree-$2s$ pseudo-expectation that satisfies $\{\|v\|^2 = 1, \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2s} \geq T_U\}$. Let $\tilde{\mathbb{E}}_U$ be the resulting pseudo-expectation for this T_U.
3. Binary search up to resolution $\sigma^2/10000$ the smallest T_L in the interval $[0, (p_{\min}^{-1} + et)^t]$ such that there exists a degree-$2t$ pseudo-expectation that satisfies $\{\|v\|^2 = 1, \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} \leq T_L\}$. Let $\tilde{\mathbb{E}}_L$ be the resulting pseudo-expectation for this T_L.
4. Return $\tilde{\mathbb{E}}_U$ and $\tilde{\mathbb{E}}_L$.

We now analyze the algorithm. To begin with, suppose that the T_U found is at least the maximum value of $\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2s}$ and that the T_L found is at most the minimum value of $\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}$. In this case $\tilde{\mathbb{E}}_U$ and $\tilde{\mathbb{E}}_L$ satisfy the axioms of Lemma 32 and Lemma 33, respectively. Then our algorithm achieves the stated guarantees:

- Suppose $\sigma^2 \geq \tau$. Note that, in this case, $M = C_{sep}k^2\sigma^2/(200e) \geq 2$. By Lemma 31, we have that
$$(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s})^{1/s} \geq \frac{C_{sep}}{100}k^2\sigma^2 \log p_{\min}^{-1} = 2eM \log p_{\min}^{-1} \geq Mes.$$
Then the conditions of Lemma 32 are satisfied, and $\tilde{\mathbb{E}}_U$ satisfies $\langle u, v\rangle^{2s} \geq (1 - 4\sigma^2/M)^s = (1-\tau)^s$.
- Suppose $\sigma^2 < \tau$ and $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \geq (4es)^s$. Note that, in this case, $M = 4$. Then the conditions of Lemma 32 are satisfied, and $\tilde{\mathbb{E}}_U$ satisfies $\langle u, v\rangle^{2s} \geq (1 - 4\sigma^2/M)^s = (1-\sigma^2)^s$.
- Suppose $\sigma^2 < 0.001$ and $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \leq (100s)^s$. Then the condition of Lemma 33 are satisfied, and $\tilde{\mathbb{E}}_L$ satisfies $\langle u, v\rangle^{2s} \geq (1 - 20\sigma^2)^s$.

We argue now that T_U is large enough and that T_L is small enough in order for the pseudo-expectations to satisfy the axioms of the lemmas. For that, we need

$$T_U \geq (1 - \sigma^2/M)\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} - \sigma^2/M\right),$$
$$T_L \leq (1 + \sigma^2/100)\left(\left(\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + \sigma^2 et\right)^t + \sigma^2/100\right).$$

We prove that the intervals in which we perform the binary search T_U and T_L contain $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s}$ and $((\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t})^{1/t} + \sigma^2 et)^t$, respectively. Then, binary search with the proposed resolutions is guaranteed to find T_U and T_L that satisfy the bounds stated above.

Using that $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^2 = 1 - \sigma^2$, we have that

$$\begin{aligned}\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t} &= \sum_{i=1}^k p_i\langle \mu_i, u\rangle^{2t} = \sum_{i=1}^k (p_i^{1/t}\langle \mu_i, u\rangle^2)^t \leq \left(\sum_{i=1}^k p_i^{1/t}\langle \mu_i, u\rangle^2\right)^t \\ &\leq \left(p_{\min}^{1/t-1}\sum_{i=1}^k p_i\langle \mu_i, u\rangle^2\right)^t = p_{\min}^{-(t-1)}\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^2 = p_{\min}^{-(t-1)}(1-\sigma^2) \\ &\leq p_{\min}^{-t}\end{aligned}$$

and

$$\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t} \geq (\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^2)^t = (1-\sigma^2)^t.$$

Therefore,

$$\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \in [(1-\sigma^2)^s, (p_{\min}^{-1})^s],$$

$$\left(\left(\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2t}\right)^{1/t} + \sigma^2 et\right)^t \in [(1-\sigma^2)^t, (p_{\min}^{-1}+et)^t].$$

Then the intervals in which we binary search are wide enough and binary search is guaranteed to succeed.

The time complexity of the algorithm is given by the number of steps in the binary search multiplied by the time to compute each of the pseudo-expectations. The number of steps in the binary search is

$$O\left(\max\left\{\log((p_{\min}^{-1})^s k^2), \log\frac{(p_{\min}^{-1})^s}{\sigma^2}, \log\frac{(p_{\min}^{-1}+et)^t}{\sigma^2}\right\}\right) = O\left(\log\frac{1}{\sigma^2} + \log^2 p_{\min}^{-1}\right).$$

For each step, we compute a pseudo-expectation of degree $O(\log p_{\min}^{-1})$ over d variables, which requires time $d^{O(\log p_{\min}^{-1})}$. Therefore the time complexity is

$$O\left(\log\frac{1}{\sigma^2} + \log^2 p_{\min}^{-1}\right)\cdot d^{O(\log p_{\min}^{-1})} = \left(\log\frac{1}{\sigma^2}\right)\cdot d^{O(\log p_{\min}^{-1})}.$$

■

C.2.2. SUM-OF-SQUARES SAMPLING (PROOF OF THEOREM 27)

We state and prove Lemma 34, which is used in the proof of Theorem 27. This lemma shows that, given a symmetric postivie definite matrix M correlated with a rank-1 matrix $uu^\top$ for a unit vector u, there exists an algorithm to recover a unit vector correlated with u. After that, we proceed to prove the theorem.

Lemma 34 (Matrix rank-1 approximation) *Let $0 \leq \epsilon < \frac{1}{8}$. Let $u \in \mathbb{R}^d$ be a unit vector. Given a symmetric positive semi-definite matrix $M \in \mathbb{R}^{d\times d}$ with $\|M\|_F \leq 1$ such that $\langle uu^\top, M\rangle_F \geq 1-\epsilon$, there exists a polynomial-time algorithm that finds a unit vector $\hat{u} \in \mathbb{R}^d$ such that $\langle u, \hat{u}\rangle^2 \geq 1-8\epsilon$.*

Proof The algorithm is to compute $vv^\top$ as the best rank-1 approximation of M and return $\frac{v}{\|v\|}$, which is uniquely defined up to a sign flip.

We now analyze the accuracy of the algorithm. We have that $\langle uu^\top, M\rangle_F \geq 1-\epsilon$, so $\|uu^\top - M\|_F^2 \leq 2 - 2\langle uu^\top, M\rangle_F \leq 2\epsilon$. For $vv^\top$ the best rank-1 approximation of M, we have then that

$$\|uu^\top - vv^\top\|_F \leq \left\|uu^\top - M\right\|_F + \left\|M - vv^\top\right\|_F \leq 2\sqrt{2\epsilon},$$

so $\|uu^\top - vv^\top\|_F^2 \le 8\epsilon$. Let $\hat{u} = \frac{v}{\|v\|}$. To analyze the error of $\hat{u}$, note that

$$\begin{aligned}\|uu^\top - vv^\top\|_F^2 &= 1 + \|v\|^4 - 2\|v\|^2 \left\langle uu^\top, \hat{u}\hat{u}^\top \right\rangle_F \\ &\ge 1 - \left\langle uu^\top, \hat{u}\hat{u}^\top \right\rangle_F \\ &= \frac{1}{2} \left\| uu^\top - \hat{u}\hat{u}^\top \right\|_F^2,\end{aligned}$$

where in the inequality we used that $1 + x^4 - 2x^2 y \ge 1 - y$ for $x \in \mathbb{R}$ and $0 \le y \le 1$, with $x = \|v\|$ and $y = \langle uu^\top, \hat{u}\hat{u}^\top \rangle$. Then $\|uu^\top - \hat{u}\hat{u}^\top\|_F^2 \le 16\epsilon$. Therefore,

$$\langle u, \hat{u} \rangle^2 = \langle uu^\top, \hat{u}\hat{u}^\top \rangle_F = 1 - \frac{1}{2}\|uu^\top - \hat{u}\hat{u}^\top\|_F^2 \ge 1 - 8\epsilon.$$

■

Proof of Theorem 27 The algorithm is to compute $M = \tilde{\mathbb{E}} vv^\top$, apply the algorithm from Lemma 34 to M in order to obtain a unit vector $\hat{u}$, and return $\hat{u}$.

We now analyze the algorithm. We start by analyzing the properties of $\tilde{\mathbb{E}}$ in more detail. Our first goal is to obtain the lower bound $\tilde{\mathbb{E}}\langle u, v \rangle^2 \ge 1 - 2\epsilon$. We start by proving the much weaker lower bound $\tilde{\mathbb{E}}\langle u, v \rangle^2 \ge 1 - t\epsilon$. Then, we use this lower bound to prove an upper bound $\tilde{\mathbb{E}}\langle u, v \rangle^{2t} \le 1 - t(1 - \tilde{\mathbb{E}}\langle u, v \rangle^2)/2$. Comparing this result with the given lower bound $\tilde{\mathbb{E}}\langle u, v \rangle^{2t} \ge (1-\epsilon)^t \ge 1 - t\epsilon$ leads to the conclusion that $\tilde{\mathbb{E}}\langle u, v \rangle^2 \ge 1 - 2\epsilon$.

We proceed with the detailed proof of this fact. Recall that $\tilde{\mathbb{E}}$ satisfies $\|v\|^2 = 1$ and $\langle u, v \rangle^{2t} \ge (1-\epsilon)^t$. We have that $\{\|v\|^2 = 1\} \left|\frac{v}{2}\right. \{0 \le \langle u, v \rangle^2 \le 1\}$, where the lower bound is trivial and the upper bound is by Lemma 57. Therefore, $\tilde{\mathbb{E}}$ also satisfies $0 \le \langle u, v \rangle^2 \le 1$.

By Lemma 58 and using that $(1-\epsilon)^t \ge 1 - t\epsilon$, we have that

$$\{0 \le \langle u, v \rangle^2 \le 1, \langle u, v \rangle^{2t} \ge (1-\epsilon)^t\} \left|\frac{v}{2t}\right. \{\langle u, v \rangle^2 \ge 1 - t\epsilon\}.$$

By Lemma 59 applied to $1 - \langle u, v \rangle^2$ with $C = \frac{1}{t^2\epsilon}$, we also have that

$$\{1 - t\epsilon \le \langle u, v \rangle^2 \le 1\} \left|\frac{v}{2t}\right. \left\{\langle u, v \rangle^{2t} \le 1 - t(1 - \langle u, v \rangle^2)/2\right\}.$$

Then

$$1 - t\epsilon \le \tilde{\mathbb{E}}\langle u, v \rangle^{2t} \le 1 - t(1 - \tilde{\mathbb{E}}\langle u, v \rangle^2)/2,$$

so by rearranging, $\tilde{\mathbb{E}}\langle u, v \rangle^2 \ge 1 - 2\epsilon$.

Then $M = \tilde{\mathbb{E}} vv^\top$ satisfies

$$\langle uu^\top, M \rangle_F = u^\top M u = \tilde{\mathbb{E}}\langle u, v \rangle^2 \ge 1 - 2\epsilon.$$

In addition, M is symmetric positive-definite and

$$\|M\|_F \le \operatorname{Tr}(M) = \tilde{\mathbb{E}} \operatorname{Tr}(vv^\top) = \tilde{\mathbb{E}}\|v\|^2 = 1.$$

Therefore, M satisfies the conditions of Lemma 34, and we are guaranteed that $\hat{u}$ satisfies $\langle u, \hat{u} \rangle^2 \ge 1 - 16\epsilon$.

The given pseudo-expectation is of degree $O(t)$ over d variables, so representing it requires $d^{O(t)}$ space. Then we simply bound the time complexity by $d^{O(t)}$, which dominates the other steps of the algorithm. ■

C.3. Finite sample bounds

In Section C.2 we assumed oracle access to $\mathbb{E}\boldsymbol{y}^{\otimes t}$. However, our algorithm only has access to empirical moments. Lemma 35 gives a sum-of-squares proof that that the empirical moments are in fact close to the population moments. We defer the proof of the lemma to the appendix.

Lemma 35 (Closeness of moments) *For $n \geq (p_{\min}^{-1} d)^{O(t)} \eta^{-2} \epsilon^{-1}$, with probability $1-\epsilon$,*

$$\left\{\|v\|^2 = 1\right\} \vdash_{O(t)}^{v} \left\{\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} \leq \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} + \eta\right\},$$

$$\left\{\|v\|^2 = 1\right\} \vdash_{O(t)}^{v} \left\{\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} \geq \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} - \eta\right\}.$$

Proof See Section F.6. ■

C.4. Proof of Theorem 22

In the setting of Theorem 22 we only have access to empirical moments. Theorem 36 and Theorem 37 adapt Theorem 26 and Theorem 25 to this setting, respectively. Also recall that the goal of Theorem 22 is to return a clustering, not only a unit vector close to u. Toward that goal, Theorem 38 shows that there exists an algorithm that, given a unit vector close to u, computes such a clustering. We state and prove all of these theorems and then combine them to prove Theorem 22.

Theorem 36 (Finite sample equivalent of Theorem 26) *Let*

$$n_0 = \left(\frac{1}{\sigma^2}\right)^{O(1)} \cdot (p_{\min}^{-1} d)^{O(\log p_{\min}^{-1})}.$$

Given a sample of size $n \geq n_0$ from the mixture, there exists an algorithm that runs in time $\left(\log \frac{1}{\sigma^2}\right) \cdot n \cdot d^{O(\log p_{\min}^{-1})}$ that computes with high probability two peseudo-expectations $\tilde{\mathbb{E}}_U$ and $\tilde{\mathbb{E}}_L$ of degree $O(\log p_{\min}^{-1})$ over a variable $v \in \mathbb{R}^d$ such that the following holds. Let $s = \lceil \log p_{\min}^{-1} \rceil$, let $t = 5000s$, and let $\tau = \frac{800e}{C_{sep} k^2}$. Then $\tilde{\mathbb{E}}_U\|v\|^2 = 1$, $\tilde{\mathbb{E}}_L\|v\|^2 = 1$, and:

- *If $\sigma^2 \geq \tau$, then $\tilde{\mathbb{E}}_U\langle u, v\rangle^{2s} \geq (1-\tau)^s$.*
- *If $\sigma^2 < \tau$ and $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \geq (4es)^s$, then $\tilde{\mathbb{E}}_U\langle u, v\rangle^{2s} \geq (1-\sigma^2)^s$.*
- *If $\sigma^2 < 0.001$ and $\mathbb{E}\langle \boldsymbol{\mu}, u\rangle^{2s} \leq (100s)^s$, then $\tilde{\mathbb{E}}_L\langle u, v\rangle^{2t} \geq (1-20\sigma^2)^t$.*

Proof The algorithm is the same as that in the proof of Theorem 26, except that in step (2) and step (3) of the algorithm the constraints that the pseudo-expectations are required to satisfy are $\{\|v\|^2 = 1, \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2s} \geq T_U\}$ and $\{\|v\|^2 = 1, \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} \leq T_L\}$, respectively.

By Lemma 35, for $n \geq n_0$, we have that with high probability

$$\{\|v\|^2 = 1, \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2s} \geq T_U\} \vdash_{2s}^{v} \{\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2s} \geq T_U - \sigma^2/(100M)\},$$

$$\{\|v\|^2 = 1, \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} \leq T_L\} \vdash_{2s}^{v} \{\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} \leq T_L + \sigma^2/10000\}.$$

These errors, combined with the errors from the binary search resolution, are still within the amount tolerated by Lemma 32 and Lemma 33, respectively, so the same guarantees hold.

The number of steps required by the binary search is the same as in Theorem 26. For each step of the binary search, we compute a pseudo-expectation of degree $O(\log p_{\min}^{-1})$ over d variables, and each constraint requires summing over the n samples, so the time required is $n \cdot d^{O(\log p_{\min}^{-1})}$. Therefore the time complexity is

$$O\left(\log \frac{1}{\sigma^2} + \log^2 p_{\min}^{-1}\right) \cdot n \cdot d^{O(\log p_{\min}^{-1})} = \left(\log \frac{1}{\sigma^2}\right) \cdot n \cdot d^{O(\log p_{\min}^{-1})}.$$

■

Theorem 37 (Finite sample equivalent of Theorem 25) *Let*

$$n_0 = \left(\frac{1}{\sigma^2}\right)^{O(1)} \cdot (p_{\min}^{-1} d)^{O(\log p_{\min}^{-1})}.$$

Given a sample of size $n \geq n_0$ from the mixture, there exists an algorithm with time complexity $\left(\log \frac{1}{\sigma^2}\right) \cdot n \cdot d^{O(\log p_{\min}^{-1})}$ that outputs with high probability a unit vector $\hat{u} \in \mathbb{R}^d$ such that $\langle u, \hat{u}\rangle^2 = 1 - 320 \min(\sigma^2, \frac{1}{k})$.

Proof The algorithm is the same as that in the proof of Theorem 25, with two exceptions:

- In step (1) of the algorithm, we run the algorithm from Theorem 36 instead of the algorithm from Theorem 26. The pseudo-expectations $\tilde{\mathbb{E}}_L$ and $\tilde{\mathbb{E}}_U$ satisfy the same guarantees.

- In step (3) of the algorithm, we check if $\hat{\mathbb{E}}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s} \geq (50s)^s$ instead of $\mathbb{E}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s} \geq (50s)^s$. By Lemma 35, for $n \geq n_0$, with high probability the difference between the two moments is less than 1. Then, if $\hat{\mathbb{E}}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s} \geq (50s)^s$, we also have $\mathbb{E}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s} \geq (50s)^s - 1$, and if $\hat{\mathbb{E}}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s} < (50s)^s$, we also have $\mathbb{E}\langle \boldsymbol{y}, \hat{u}_U\rangle^{2s} < (50s)^s + 1$. It is easy to verify that the analysis in Theorem 25 is still correct with these slightly weaker bounds.

Therefore, the same guarantees hold as in Theorem 25. The time complexity of the algorithm is dominated by the time to run the algorithm from Theorem 36.

■

Theorem 38 (Clustering algorithm) *For some $C > 0$, suppose that a unit vector $\hat{u} \in \mathbb{R}^d$ is known such that $\langle u, \hat{u}\rangle^2 \geq 1 - C \min(\sigma^2, \frac{1}{k})$. Suppose that C_{sep}/C is larger than some universal constant. Then, given a sample of size $n \geq (p_{\min}^{-1})^{O(1)}$ from the mixture, there exists an algorithm that runs in time $n^{O(\log p_{\min}^{-1})}$ and returns with high probability a partition of $[n]$ into k sets $C_1, ..., C_k$ such that, if the true clustering of the samples is $S_1, ..., S_k$, then there exists a permutation π of $[k]$ such that*

$$1 - \frac{1}{n} \sum_{i=1}^{k} |C_i \cap S_{\pi(i)}| \leq \left(\frac{p_{\min}}{k}\right)^{O(1)}.$$

Proof Our algorithm runs the algorithm from Theorem 5.1 of Hopkins and Li (2018) with some $t = O(\log p_{\min}^{-1})$ large enough on input samples $\langle y_1, \hat{u}\rangle/(\sqrt{2(C+1)\sigma^2})$, ..., $\langle y_n, \hat{u}\rangle/(\sqrt{2(C+1)\sigma^2})$, and returns the clustering that this algorithm computes as an intermediate step.

We now analyze the algorithm. Note that $\langle \boldsymbol{y}, \hat{u}\rangle/\sqrt{2(C+1)\sigma^2}$ is distributed according to a one-dimensional mixture of Gaussians in which all the components have the same variance, namely $\hat{u}^\top \Sigma \hat{u}/(2(C+1)\sigma^2)$. We have that

$$\hat{u}^\top \Sigma \hat{u} = 1 - (1-\sigma^2)\langle u, \hat{u}\rangle^2 = 1 - \langle u, \hat{u}\rangle^2 + \sigma^2 \langle u, \hat{u}\rangle^2 \le C\sigma^2 + \sigma^2 = (C+1)\sigma^2.$$

Therefore, the variance $\hat{u}^\top \Sigma \hat{u}/(2(C+1)\sigma^2)$ is upper bounded by $1/2$. For the guarantees of the algorithm from Hopkins and Li (2018) to hold, we further need to show that the mixture has large separation between the means of the components. Note that the mean corresponding to μ_i in the original mixture becomes $\langle \mu_i, \hat{u}\rangle/\sqrt{2(C+1)\sigma^2}$ in the new mixture. For $i \ne j$, we have that

$$\begin{aligned}(\langle \mu_i - \mu_j, \hat{u}\rangle)^2 &= \langle \langle \mu_i, u\rangle u - \langle \mu_j, u\rangle u, \hat{u}\rangle^2 = \langle u, \hat{u}\rangle^2 \langle \mu_i - \mu_j, u\rangle^2 \\ &\ge \langle u, \hat{u}\rangle^2 C_{sep}(u^\top \Sigma u) \log p_{\min}^{-1} = \langle u, \hat{u}\rangle^2 C_{sep}\sigma^2 \log p_{\min}^{-1} \\ &\ge \frac{C_{sep}}{2}\sigma^2 \log p_{\min}^{-1}\end{aligned}$$

where in the last inequality we used that $\langle u, \hat{u}\rangle^2 \ge 1 - C/k \ge 1/2$. Then

$$\left(\frac{\langle \mu_i - \mu_j, \hat{u}\rangle}{\sqrt{2(C+1)\sigma^2}}\right)^2 \ge \frac{C_{sep}}{4(C+1)} \log p_{\min}^{-1}.$$

For C_{sep}/C larger than some universal constant, the separation coefficient $C_{sep}/(4(C+1))$ is large enough for the guarantees of Theorem 5.1 of Hopkins and Li (2018) to hold meaningfully with $t = O(\log p_{\min}^{-1})$. Then this algorithm computes a clustering with the stated guarantees. The algorithm requries $n \ge (p_{\min}^{-1})^{O(1)}$ and runs in time $n^{O(\log p_{\min}^{-1})}$. ■

Proof of Theorem 22 Run the algorithm from Theorem 37 to obtain a unit vector $\hat{u}$ that satisfies $\langle u, \hat{u}\rangle^2 \ge 1 - 320\min(\sigma^2, \frac{1}{k})$. Then run the clustering algorithm from Theorem 38 using this unit vector $\hat{u}$. For $C_{sep}/320$ larger than some universal constant, this algorithm is guaranteed to return a clustering with the stated guarantees.

The time complexity from Theorem 37 is $\left(\log \frac{1}{\sigma^2}\right) \cdot n \cdot d^{O(\log p_{\min}^{-1})}$ and the time complexity from Theorem 39 is $n^{O(\log p_{\min}^{-1})}$. We assume $n \ge \left(\frac{1}{\sigma^2}\right)^{O(1)} \cdot (p_{\min}^{-1} d)^{O(\log p_{\min}^{-1})}$. Therefore, the time complexity is dominated by the time to run the clustering algorithm from Theorem 39. ■

Appendix D. Colinear means

In this section we remove the isotropic position assumption from the model in Section C. Our strategy is straightfoward: we first put the mixture in isotropic position and then run the algorithm from Theorem 22. The technical challenge is that we can only put the mixture in approximate isotropic position. Then, we show that the guarantees of Theorem 22 continue to hold with approxiamte isotropic position, albeit with a sample complexity that depends on the condition number of the covariance matrix of the mixture.

Setting. We consider a mixture of k Gaussian distributions $N(\mu_i^0, \Sigma^0)$ with mixing weights p_i for $i = 1, ..., k$, where $\mu_i^0 \in \mathbb{R}^d$, $\Sigma^0 \in \mathbb{R}^{d \times d}$ is positive definite, and $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$. Let $p_{\min} = \min_i p_i$.

The distribution also satisfies mean separation and mean colinearity:

- Mean separation: for some $C_{sep} > 0$ and for all $i \neq j$,

$$\left\|\left(\Sigma^0\right)^{-1/2}\left(\mu_i^0 - \mu_j^0\right)\right\|^2 \geq C_{sep} \log p_{\min}^{-1}.$$

- Mean colinearity: for some vector $\mu_{base}^0 \in \mathbb{R}^d$ and some unit vector $u^0 \in \mathbb{R}^d$ and for all i,

$$\mu_i^0 = \mu_{base}^0 + \langle \mu_i^0 - \mu_{base}^0, u^0 \rangle u^0.$$

Also define, for $\boldsymbol{y}^0$ distributed according to the mixture,

$$\sigma^2 = \frac{(u^0)^\top \operatorname{cov}(\boldsymbol{y}^0)^{-1} u^0}{(u^0)^\top (\Sigma^0)^{-1} u^0}.$$

As shown in Lemma 41, σ^2 has the same interpretation as in Section C: it is equal to the variance of the components in the direction of the means after an isotropic position transformation.

Theorem 39 (Colinear means algorithm) *Consider the Gaussian mixture model defined above, with C_{sep} larger than some universal constant. For $\boldsymbol{y}^0$ distributed according to the mixture, let*

$$n_0 = \left(\frac{1}{\sigma^2} \cdot \|\operatorname{cov}(\boldsymbol{y}^0)\| \cdot \|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|\right)^{O(1)} \cdot (p_{\min}^{-1} d)^{O(\log p_{\min}^{-1})}.$$

Given a sample of size $n \geq n_0$ from the mixture, there exists an algorithm that runs in time $n^{O(\log p_{\min}^{-1})}$ and returns with high probability a partition of $[n]$ into k sets $C_1, ..., C_k$ such that, if the true clustering of the samples is $S_1, ..., S_k$, then there exists a permutation π of $[k]$ such that

$$1 - \frac{1}{n} \sum_{i=1}^k |C_i \cap S_{\pi(i)}| \leq \left(\frac{p_{\min}}{k}\right)^{O(1)}.$$

We introduce some further notation for this section. Let $\boldsymbol{y}^0$ be distributed according to the mixture. We specify the model as $\boldsymbol{y}^0 = \boldsymbol{\mu}^0 + \boldsymbol{w}^0$, where $\boldsymbol{\mu}^0$ takes value μ_i with probability p_i and $\boldsymbol{w}^0 \sim N(0, \Sigma^0)$, with $\boldsymbol{\mu}^0$ and $\boldsymbol{w}^0$ independent of each other.

D.1. Isotropic position transformation

In this section we argue that, if we put the mixture in exact isotropic position, the conditions of Theorem 22 are satisfied and we can simply run that algorithm.

Assume acces to $\mathbb{E}\boldsymbol{y}^0$ and to an invertible matrix $W \in \mathbb{R}^{d \times d}$ such that $W \operatorname{cov}(\boldsymbol{y}^0) W^\top = I_d$. Then, define the random variable $\boldsymbol{y}$ by the affine transformation $\boldsymbol{y} = W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0)$. The distribution of $\boldsymbol{y}$ is in isotropic position: it has mean 0 and covariance matrix I_d. Furthermore, Lemma 40 shows that $\boldsymbol{y}$ is a mixture of Gaussians in which the components are affine transformed versions of the original components, and that the mixture continues to satisfy mean separation and mean colinearity. Then, the conditions of Theorem 22 are satisfied. Therefore, if the original input samples are $y_1^0, ..., y_n^0$, we can simply run that algorithm on input samples $W(y_1^0 - \mathbb{E}y^0)$, ..., $W(y_n^0 - \mathbb{E}y^0)$[15].

We define now some variables used to state Lemma 40. Recall that $\boldsymbol{y} = W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0)$. Define $\mu_i = W(\mu_i^0 - \mathbb{E}\boldsymbol{y}^0)$, $\Sigma = W\Sigma^0 W^\top$, and $u = \frac{Wu^0}{\|Wu^0\|v}$. Also define random variables $\boldsymbol{\mu} = W(\boldsymbol{\mu}^0 - \mathbb{E}\boldsymbol{y}^0)$ and $\boldsymbol{w} = W\boldsymbol{w}^0$.

Lemma 40 (Model after isotropic position transformation) *The random variable $\boldsymbol{y}$ is distributed according to a mixture of k Gaussian distributions $N(\mu_i, \Sigma)$ with mixing weights p_i for $i = 1, ..., k$, where Σ is positive definite. Alternatively, we specify the model as $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{w}$, with $\boldsymbol{\mu}$ and $\boldsymbol{w}$ independent of each other. Furthermore, for all $i \neq j$ we have mean separation*

$$\left\|\Sigma^{-1/2}(\mu_i - \mu_j)\right\|^2 \geq C_{sep} \log p_{\min}^{-1}$$

and for all i we have mean colinearity

$$\mu_i = \langle \mu_i, u \rangle u.$$

Proof Recall that $\boldsymbol{y}^0 = \boldsymbol{\mu}^0 + \boldsymbol{w}^0$. Then $W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0) = W(\boldsymbol{\mu}^0 - \mathbb{E}\boldsymbol{y}^0) + W\boldsymbol{w}^0$, so $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{w}$. Also note that $\boldsymbol{\mu}$ takes value $W(\mu_i^0 - \mathbb{E}\boldsymbol{y}^0) = \mu_i$ with probability p_i and $\boldsymbol{w} \sim N(0, W\Sigma^0 W^\top) = N(0, \Sigma)$. Therefore, $\boldsymbol{y}$ is distributed according to a mixture of k Gaussian distributions $N(\mu_i, \Sigma)$ with mixing weights p_i.

To show that Σ is positive definite, we note that for any vector $v \in \mathbb{R}^d$ with $v \neq 0$ we have that

$$v^\top \Sigma v = v^\top W \Sigma^0 W^\top v = (W^\top v)^\top \Sigma^0 W^\top v > 0,$$

where we used that $W^\top v \neq 0$ because W is invertible, after which we used that Σ^0 is positive definite.

We prove now mean colinearity and mean separation. We start with mean colinearity. Recall that $\mu_i^0 = \mu_{base}^0 + \langle \mu_i^0 - \mu_{base}^0, u^0 \rangle u^0$. Then, using that $\mathbb{E}\boldsymbol{y}^0 = \mathbb{E}\boldsymbol{\mu}^0$,

$$\mathbb{E}\boldsymbol{y}^0 = \mu_{base}^0 + \langle \mathbb{E}\boldsymbol{y}^0 - \mu_{base}^0, u^0 \rangle u^0,$$

so

$$\mu_i = W(\mu_i^0 - \mathbb{E}\boldsymbol{y}^0) = W(\langle \mu_i^0 - \mathbb{E}\boldsymbol{y}^0, u^0 \rangle u^0) = \langle \mu_i^0 - \mathbb{E}\boldsymbol{y}^0, u^0 \rangle W u^0.$$

Then, using that $u = \frac{Wu^0}{\|Wu^0\|}$,

$$\begin{aligned}
\langle \mu_i, u \rangle u &= \left\langle \langle \mu_i^0 - \mathbb{E}\boldsymbol{y}^0, u^0 \rangle W u^0, \frac{Wu^0}{\|Wu^0\|} \right\rangle \frac{Wu^0}{\|Wu^0\|} \\
&= \langle \mu_i^0 - \mathbb{E}\boldsymbol{y}^0, u^0 \rangle \left\langle \frac{Wu^0}{\|Wu^0\|}, \frac{Wu^0}{\|Wu^0\|} \right\rangle Wu^0 \\
&= \langle \mu_i^0 - \mathbb{E}\boldsymbol{y}^0, u^0 \rangle W u^0 \\
&= \mu_i,
\end{aligned}$$

[15] It is straightforward that if sample y_i^0 comes from the i-th component in the original mixture then $W(y_i^0 - \mathbb{E}y^0)$ continues to come from the i-th component in the affine transformed mixture.

which proves mean colinearity. For mean separation, we have that

$$\begin{aligned}\left\|\Sigma^{-1/2}(\mu_i-\mu_j)\right\|^2 &= \left\|(W\Sigma^0W^\top)^{-1/2}W(\mu_i^0-\mu_j^0)\right\|^2\\ &= (\mu_i^0-\mu_j^0)^\top W^\top(W\Sigma^0W^\top)^{-1}W(\mu_i^0-\mu_j^0)\\ &= (\mu_i^0-\mu_j^0)^\top W^\top(W^\top)^{-1}(\Sigma^0)^{-1}W^{-1}W(\mu_i^0-\mu_j^0))\\ &= (\mu_i^0-\mu_j^0)^\top(\Sigma^0)^{-1}(\mu_i^0-\mu_j^0)\\ &= \left\|(\Sigma^0)^{-1/2}(\mu_i^0-\mu_j^0)\right\|^2\\ &\geq C_{sep}\log p_{\min}^{-1}.\end{aligned}$$

■

Lemma 41

$$\sigma^2 = \frac{(u^0)^\top \operatorname{cov}(\boldsymbol{y}^0)^{-1}u^0}{(u^0)^\top(\Sigma^0)^{-1}u^0} = u^\top\Sigma u.$$

Proof For the purposes of this proof, we define $\sigma^2 = u^\top\Sigma u$ as in Section C.1 and prove that it also matches the definition in this section.

We have, as in the proof of Lemma 40, that

$$\|(\Sigma^0)^{-1/2}(\mu_i^0-\mu_j^0)\|^2 = \|\Sigma^{-1/2}(\mu_i-\mu_j)\|^2.$$

For the left-hand side, we have that

$$\begin{aligned}\|(\Sigma^0)^{-1/2}(\mu_i^0-\mu_j^0)\|^2 &= \|(\Sigma^0)^{-1/2}\langle\mu_i^0-\mu_j^0,u^0\rangle u^0\|^2\\ &= \|(\Sigma^0)^{-1/2}u^0\|^2\cdot\langle\mu_i^0-\mu_j^0,u^0\rangle^2.\end{aligned}$$

For the right-hand side, using from Lemma 23 that $\Sigma = I_d-(1-\sigma^2)uu^\top$, we have that

$$\begin{aligned}\|\Sigma^{-1/2}(\mu_i-\mu_j)\|^2 &= \left\|\left(I_d-(1-\sigma^2)uu^\top\right)^{-1/2}(\mu_i-\mu_j))\right\|^2\\ &= \left\|\left(I_d-(1-\sigma^2)\frac{(Wu^0)(Wu^0)^\top}{\|Wu^0\|^2}\right)^{-1/2}\langle\mu_i^0-\mu_j^0,u^0\rangle Wu^0\right\|^2\\ &= \left\|\left(I_d+\left(\frac{1}{\sigma}-1\right)\frac{(Wu^0)(Wu^0)^\top}{\|Wu^0\|^2}\right)Wu^0\right\|^2\cdot\langle\mu_i^0-\mu_j^0,u^0\rangle^2\\ &= \frac{1}{\sigma^2}\cdot\|Wu^0\|^2\cdot\langle\mu_i^0-\mu_j^0,u^0\rangle^2.\end{aligned}$$

Therefore

$$\|(\Sigma^0)^{-1/2}u^0\|^2\cdot\langle\mu_i^0-\mu_j^0,u^0\rangle^2 = \frac{1}{\sigma^2}\cdot\|Wu^0\|^2\cdot\langle\mu_i^0-\mu_j^0,u^0\rangle^2,$$

so

$$\sigma^2 = \frac{\|Wu^0\|^2}{\|(\Sigma^0)^{-1/2}u^0\|^2} = \frac{(u^0)^\top\operatorname{cov}(\boldsymbol{y}^0)^{-1}u^0}{(u^0)^\top(\Sigma^0)^{-1}u^0},$$

where we used that, by Lemma 74, $W = Q\operatorname{cov}(\boldsymbol{y}^0)^{-1/2}$ for an orthogonal matrix Q, so $\|Wu^0\| = \|\operatorname{cov}(\boldsymbol{y}^0)^{-1/2}u^0\|$. ■

D.2. Finite sample isotropic position transformation

Without access to $\mathbb{E}\boldsymbol{y}^0$ and to W, we apply the isotropic position transformation with $\hat{\mathbb{E}}\boldsymbol{y}^0$ and some matrix $\hat{W} \in \mathbb{R}^{d\times d}$ defined as follows. Let the singular value decomposition of $\widehat{\operatorname{cov}}(\boldsymbol{y}^0)$ be $\hat{U}\hat{\Lambda}\hat{U}^\top$. Then define $\hat{W}$ and W as

$$\hat{W} = (\hat{U}^\top \widehat{\operatorname{cov}}(\boldsymbol{y}^0)\hat{U})^{-1/2}\hat{U}^\top, \quad W = (\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{-1/2}\hat{W}. \tag{19}$$

This choice is analogous to that in Appendix C in Hsu and Kakade (2013). By Lemma 72, we have that $\hat{W}\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\hat{W}^\top = I_d$ and $W\operatorname{cov}(\boldsymbol{y}^0)W^\top = I_d$. Hence, $\hat{W}$ corresponds to an isotropic position transformation for the empirical covariance matrix, and W to one for the population covariance matrix. In our algorithm, we will apply the approximate isotropic position transformation to input samples $y_1^0, ..., y_n^0$ as $\hat{W}(y_1^0 - \hat{\mathbb{E}}y^0), ..., \hat{W}(y_n^0 - \hat{\mathbb{E}}y^0)$.

D.3. Finite sample bounds

Lemma 42 gives a sum-of-squares proof that the empirical moments of the mixture with approximate isotropic position transformation are close to the population moments of the mixture with exact isotropic position transformation. This lemma is supported by Lemma 43 and Lemma 44, which prove that the moments do not change much due to the use of $\hat{\mathbb{E}}\boldsymbol{y}^0$ and $\hat{W}$, respectively.

Additionally, for arbitrary unit vectors v, Lemma 45 proves that $\langle \hat{W}(\mu_i^0 - \mu_j^0), v\rangle$ is close to $\langle W(\mu_i^0 - \mu_j^0), v\rangle$ and Lemma 46 proves that $v^\top \hat{W}(\Sigma^0)^{1/2}$ is close to $v^\top W(\Sigma^0)^{1/2}$. These facts are used in the proof of Theorem 39 to argue that the clustering algorithm is correct.

Lemma 42 (Closeness of moments) *Let* $\eta < 0.001$. *For*

$$n \geq \left(\|\operatorname{cov}(\boldsymbol{y}^0)\| \cdot \|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|\right)^{O(1)} \cdot (p_{\min}^{-1} d)^{O(t)} \eta^{-2}\epsilon^{-1},$$

with probability $1-\epsilon$,

$$\left\{\|v\|^2 = 1\right\} \Big|\frac{v}{2t} \left\{ \hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} \leq (1+\eta)\cdot \mathbb{E}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \eta \right\},$$

$$\left\{\|v\|^2 = 1\right\} \Big|\frac{v}{2t} \left\{ \hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} \geq (1-\eta)\cdot \mathbb{E}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} - \eta \right\}.$$

Proof Select n such that the results of Lemma 43, Lemma 44, and Lemma 35 hold each with probability $1-\epsilon/3$. Then, overall, all three results hold with probability $1-\epsilon$. Then we have with probability $1-\epsilon$ that

$$\begin{aligned}
&\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} \\
&\quad\leq (1+\eta/10)\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \eta/10 \\
&\quad\leq (1+\eta/10)\left((1+\eta/10)\hat{\mathbb{E}}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \eta/10\right) + \eta/10 \\
&\quad\leq (1+\eta/10)\left((1+\eta/10)\left((1+\eta/10)\mathbb{E}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \eta/10\right) + \eta/10\right) + \eta/10 \\
&\quad\leq (1+\eta)\mathbb{E}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \eta.
\end{aligned}$$

and

$$\begin{aligned}
&\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} \\
&\quad \geq (1-\eta/10)\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} - \eta/10 \\
&\quad \geq (1-\eta/10)\left((1-\eta/10)\hat{\mathbb{E}}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} - \eta/10\right) - \eta/10 \\
&\quad \geq (1-\eta/10)\left((1-\eta/10)\left((1-\eta/10)\mathbb{E}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} - \eta/10\right) - \eta/10\right) - \eta/10 \\
&\quad \geq (1-\eta)\mathbb{E}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} - \eta.
\end{aligned}$$

■

Lemma 43 *Let* $\eta < 0.001$. *For*

$$n \geq kd\log^2(d/\epsilon)\cdot\left(\frac{t\cdot\|\operatorname{cov}(\boldsymbol{y}^0)\|\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|}{\eta}\right)^{O(1)},$$

with probability $1-\epsilon$,

$$\left\{\|v\|^2 = 1\right\} \left|\frac{v}{2t}\right. \left\{\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} \leq (1+\eta)\cdot\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \eta\right\},$$

$$\left\{\|v\|^2 = 1\right\} \left|\frac{v}{2t}\right. \left\{\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} \geq (1-\eta)\cdot\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} - \eta\right\}.$$

Proof See Section F.7. ■

Lemma 44 *Let* $\eta < 0.001$. *For*

$$n \geq kd\log^2(d)\epsilon^{-1}\cdot\left(\frac{tp_{\min}^{-1}d\cdot\|\operatorname{cov}(\boldsymbol{y}^0)\|\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|}{\eta}\right)^{O(1)},$$

with probability $1-\epsilon$,

$$\left\{\|v\|^2 = 1\right\} \left|\frac{v}{2t}\right. \left\{\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} \leq (1+\eta)\cdot\hat{\mathbb{E}}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \eta\right\},$$

$$\left\{\|v\|^2 = 1\right\} \left|\frac{v}{2t}\right. \left\{\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} \geq (1-\eta)\cdot\hat{\mathbb{E}}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} - \eta\right\}.$$

Proof See Section F.7. ■

Lemma 45 *Let* $\eta < 0.001$. *Let* $v \in \mathbb{R}^d$ *be a unit vector. For*

$$n \geq kd\log^2(d/\epsilon)\cdot\left(\frac{\|\operatorname{cov}(\boldsymbol{y}^0)\|\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|}{\eta}\right)^{O(1)},$$

with probability $1-\epsilon$, *for all* $i, j \in [k]$,

$$|\langle W(\mu_i^0 - \mu_j^0), v\rangle - \langle \hat{W}(\mu_i^0 - \mu_j^0), v\rangle| \leq \eta.$$

Proof See Section F.7. ■

Lemma 46 *Let $\eta < 0.001$. Let $v \in \mathbb{R}^d$ be a unit vector. For*

$$n \geq kd\log^2(d/\epsilon) \cdot \left(\frac{p_{\min}^{-1} \cdot \|\operatorname{cov}(\boldsymbol{y}^0)\| \cdot \|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|}{\eta}\right)^{O(1)},$$

with probability $1-\epsilon$,

$$\|v^\top W(\Sigma^0)^{1/2} - v^\top \hat{W}(\Sigma^0)^{1/2}\| \leq \eta.$$

Proof See Section F.7. ■

D.4. Proof of Theorem 39

Proof of Theorem 39 The first step of the algorithm is to apply the approximate istropic position transformation described in Section D.2 to input samples $y_1^0, ..., y_n^0$. Then, the new samples are $\hat{W}(y_1^0 - \hat{\mathbb{E}}y^0), ..., \hat{W}(y_n^0 - \hat{\mathbb{E}}y^0)$. After that, the algorithm is the same as that of Theorem 22.

We now argue that, for $n \geq n_0$, the same guarantees as in Theorem 22 hold. Recall that Theorem 22 is composed of two parts: the algorithm of Theorem 37 which computes a unit vector $\hat{u}$ with correlation $1 - 320\min(\sigma^2, \frac{1}{k})$ with u, and the clustering algorithm of Theorem 38, which uses a unit vector $\hat{u}$ with such correlation in order to cluster the samples.

For the algorithm of Theorem 37, we note that by Lemma 42, for $n \geq n_0$, we have sum-of-squares proofs that, for $t = O(\log p_{\min}^{-1})$,

$$\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} \leq \left(1 + \frac{\sigma^2}{10000M}\right)\mathbb{E}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \frac{\sigma^2}{10000M}$$

and

$$\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} \geq \left(1 - \frac{\sigma^2}{10000M}\right)\mathbb{E}\langle W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} - \frac{\sigma^2}{10000M},$$

where $\hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0)$ corresponds to the mixture in approximate isotropic position and $W(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0)$ corresponds to the mixture in exact isotropic position. It is easy to verify, similarly to the analysis of the errors in the proof of Theorem 37, that these errors are tolerated by the algorithm and that it behaves as if the distribution was in exact isotropic position.

For the clustering algorithm of Theorem 38, the main issue is that the samples are colinear in direction $\frac{\hat{W}u^0}{\|\hat{W}u^0\|}$, but $\hat{u}$ is guaranteed to have large correlation with $\frac{Wu^0}{\|Wu^0\|}$. We prove, nevertheless, that the algorithm has the same guarantees. First, we show that the variance of the one-dimensional components $\hat{u}^\top \hat{W}^\top \Sigma^0 \hat{W}\hat{u}/(2(C+1)\sigma^2)$ is upper bounded by 1, as required by the algorithm of Hopkins and Li (2018). We have by Lemma 46 that, for $n \geq n_0$, with high probability

$$|\|\hat{u}^\top W(\Sigma^0)^{1/2}\| - \|\hat{u}^\top \hat{W}(\Sigma^0)^{1/2}\|| \leq \|\hat{u}^\top W(\Sigma^0)^{1/2} - \hat{u}^\top \hat{W}(\Sigma^0)^{1/2}\| \leq \sigma/100,$$

so using that $\hat{u}^\top W\Sigma^0 W^T\hat{u} \geq \sigma^2$,

$$\hat{u}^\top W\Sigma^0 W^T \hat{u} \geq \frac{1}{2}\hat{u}^\top \hat{W}\Sigma^0 \hat{W}^T \hat{u}.$$

Therefore, using from the proof of Theorem 38 that $(\hat{u}^\top W^\top \Sigma^0 W \hat{u})/(2(C+1)\sigma^2) \leq 0.5$,

$$\frac{\hat{u}^\top \hat{W}^\top \Sigma^0 \hat{W} \hat{u}}{2(C+1)\sigma^2} \leq 2 \frac{\hat{u}^\top W^\top \Sigma^0 W \hat{u}}{2(C+1)\sigma^2} \leq 1.$$

Second, we show that the one-dimensional means $\langle \hat{W}(\mu_i^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), \hat{u} \rangle$ have large separation. We have by the proof of Theorem 38 that, for $n \geq n_0$, with high probability

$$\langle W(\mu_i^0 - \mu_j^0), \hat{u} \rangle^2 \geq \frac{C_{sep}}{2} \sigma^2 \log p_{\min}^{-1}.$$

We are interested in a similar bound with W changed into $\hat{W}$. By Lemma 45, for $n \geq n_0$, we have with high probability

$$|\langle W(\mu_i^0 - \mu_j^0), \hat{u} \rangle - \langle \hat{W}(\mu_i^0 - \mu_j^0), \hat{u} \rangle| \leq \sigma/100,$$

so using that $\langle W(\mu_i^0 - \mu_j^0), \hat{u} \rangle^2 \geq \sigma^2$,

$$\langle \hat{W}(\mu_i^0 - \mu_j^0), \hat{u} \rangle^2 \geq \frac{1}{2} \langle W(\mu_i^0 - \mu_j^0), \hat{u} \rangle^2.$$

Therefore,

$$\langle \hat{W}(\mu_i^0 - \mu_j^0), \hat{u} \rangle^2 \geq \frac{C_{sep}}{4} \sigma^2 \log p_{\min}^{-1},$$

so

$$\left(\frac{\langle \hat{W}(\mu_i^0 - \mu_j^0), \hat{u} \rangle}{\sqrt{2(C+1)\sigma^2}} \right)^2 \geq \frac{C_{sep}}{8(C+1)} \log p_{\min}^{-1}.$$

For C_{sep}/C larger than some universal constant, the separation coefficient $C_{sep}/(8(C+1))$ is large enough for the guarantees of Theorem 5.1 of Hopkins and Li (2018) to hold as before with $t = O(\log p_{\min}^{-1})$.

Then, overall, the same guarantees as in Theorem 22 hold. ■

Appendix E. Small radius

Setting. We consider a mixture of k Gaussian distributions $N(\mu_i, \Sigma)$ with mixing weights p_i for $i = 1, ..., k$, where $\mu_i \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite, and $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$. Let $p_{\min} = \min_i p_i$.

The distribution also satisfies mean separation and a small radius condition:

- Mean separation: for some $C_{sep} > 0$ and for all $i \neq j$,

$$\left\| \Sigma^{-1/2}(\mu_i - \mu_j) \right\|^2 \geq C_{sep} \log p_{\min}^{-1}.$$

- Small radius: for some $R > 0$ and for all i,

$$\left\| \Sigma^{-1/2} \mu_i \right\| \leq R.$$

Theorem 47 (Small radius algorithm) *Consider the Gaussian mixture model defined above, with C_{sep} larger than some universal constant. Let $n_0 = (p_{\min}^{-1} d)^{O(R^2 + \log p_{\min}^{-1})}$. Given a sample of size $n \geq n_0$ from the mixture, there exists an algorithm that runs in time $n^{O(R^2 + \log p_{\min}^{-1})}$ and returns with high probability a partition of $[n]$ into k sets $C_1, ..., C_k$ such that, if the true clustering of the samples is $S_1, ..., S_k$, then there exists a permutation π of $[k]$ such that*

$$1 - \frac{1}{n} \sum_{i=1}^{k} |C_i \cap S_{\pi(i)}| \leq \left(\frac{p_{\min}}{k} \right)^{O(1)}.$$

We introduce some further notation for this section. Let $\boldsymbol{y}$ be distributed according to the mixture. We specify the model as $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{w}$, where $\boldsymbol{\mu}$ takes value μ_i with probability p_i and $\boldsymbol{w} \sim N(0, \Sigma)$, with $\boldsymbol{\mu}$ and $\boldsymbol{w}$ independent of each other.

E.1. Component covariance estimation

Lemma 49 gives a sum-of-squares proof that, for $t = \Omega(R^4)$, the directional moment $\mathbb{E}\langle \boldsymbol{y}, v \rangle^{2t}$ approximates the t-th power of the variance of the components in direction v. This is the main ingredient of the algorithm, and it shows that the t-th moment of the distribution identifies within constant factors the covariance matrix of the components. Lemma 48 is a simple upper bound on the means of the mixture, used in the proof of Lemma 49.

Lemma 48 (Bounded mean term) *For $t \geq 1$ integer,*

$$\Big|\frac{v}{2t}\; \mathbb{E}\langle \boldsymbol{\mu}, v \rangle^{2t} \leq R^{2t} (v^\top \Sigma v)^t.$$

Proof

$$\Big|\frac{v}{2t}\; \mathbb{E}\langle \boldsymbol{\mu}, v \rangle^{2t} = \mathbb{E}\langle \Sigma^{1/2}\Sigma^{-1/2}\boldsymbol{\mu}, v \rangle^{2t} = \mathbb{E}\langle \Sigma^{-1/2}\boldsymbol{\mu}, \Sigma^{1/2} v \rangle^{2t} \leq \mathbb{E}\|\Sigma^{-1/2}\boldsymbol{\mu}\|^{2t} \|\Sigma^{1/2} v\|^{2t},$$

where in the inequality we used Lemma 57. The conclusion follows by noting that $\|\Sigma^{-1/2}\boldsymbol{\mu}\| \leq R$ and $\|\Sigma^{1/2} v\|^2 = v^\top \Sigma v$. ■

Lemma 49 (Small radius component covariance estimation) *For $t \geq 4R^2$ integer,*

$$\Big|\frac{v}{2t}\; \frac{1}{4^t} (v^\top \Sigma v)^t \leq \frac{1}{t^t} \mathbb{E}\langle \boldsymbol{y}, v \rangle^{2t} \leq 4^t (v^\top \Sigma v)^t.$$

Proof For the upper bound, by Lemma 13 and Lemma 48,

$$\begin{aligned} \Big|\frac{v}{2t}\; \mathbb{E}\langle \boldsymbol{y}, v \rangle^{2t} &\leq 2^{2t-1} \mathbb{E}\langle \boldsymbol{\mu}, v \rangle^{2t} + 2^{2t-1} (v^\top \Sigma v)^t t^t \\ &\leq 2^{2t-1} (R^{2t} + t^t)(v^\top \Sigma v)^t \\ &\leq 4^t t^t (v^\top \Sigma v)^t. \end{aligned}$$

For the lower bound, by Lemma 14 and Lemma 48,

$$\begin{aligned} \Big|\frac{v}{2t}\; \mathbb{E}\langle \boldsymbol{y}, v \rangle^{2t} &\geq \mathbb{E}\langle \boldsymbol{\mu}, v \rangle^{2t} + (v^\top \Sigma v)^t \frac{t^t}{2^t} \\ &\geq \left(-R^{2t} + \frac{1}{2^t} t^t \right) (v^\top \Sigma v)^t \\ &\geq \frac{1}{4^t} t^t (v^\top \Sigma v)^t. \end{aligned}$$

■

We note that the term $\frac{1}{t^t}\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}$ can be rewritten as

$$\frac{1}{t^t}\mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} = \left\langle v^{\otimes t}, \frac{1}{t^t}\mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} v^{\otimes t}\right\rangle = \left\langle \frac{1}{t^t}\mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}, (vv^\top)^{\otimes t}\right\rangle.$$

E.2. Finite sample bounds

Lemma 50 gives a sum-of-squares proof that the empirical moments of the distribution are close to the population moments. We defer the proof to the appendix.

Lemma 50 (Closeness of moments) *For $n \geq (p_{\min}^{-1} d)^{O(t)} \eta^{-2}\epsilon^{-1}$, with probability $1-\epsilon$,*

$$\Big|\frac{v}{2t}\ (1-\eta)\cdot \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} \leq \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} \leq (1+\eta)\cdot \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}.$$

Proof See Section F.8. ■

E.3. Proof of Theorem 47

If the covariance matrix of the components were known, we could apply an affine transformation to the samples and change the distribution into a mixture of spherical Gaussians. After that we could simply apply an algorithm for clustering mixtures of spherical Gaussians.

It might look like the covariance matrix approximation of Lemma 49 could be used to design a sum-of-squares program that identifies this covariance matrix. However, because Lemma 49 only gives an approximation in each direction for the t-th power of the variance of the components, and because it is non-trivial to take t-th roots in sum-of-squares proofs, we found it challenging to obtain a low-degree sum-of-squares proof of identifiability for the covariance matrix.

Instead, we observe that the sum-of-squares algorithm of Hopkins and Li (2018) for clustering mixtures of spherical Gaussians only uses as axioms upper bounds on the t-th moments of the distribution of the components. It is not difficult to adapt this algorithm to work with the t-th power approximations that we obtain from Lemma 49.

Proof of Theorem 47 The algorithm is:

1. Set $t = O(R^2 + \log p_{\min}^{-1})$ large enough.
2. Estimate $D = \frac{1}{t^t}\hat{\mathbb{E}}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}$.
3. Apply the algorithm from Theorem 5.1 of Hopkins and Li (2018), but with the following moment constraint in the set of axioms instead of the original moment constraint:

$$\forall v \in \mathbb{R}^d, \frac{1}{\alpha n}\sum_{i=1}^{n} w_i \langle y_i - \mu, v\rangle^{2t} \leq 4(8t)^t \langle v^{\otimes t}, D v^{\otimes t}\rangle,$$

where α is a parameter and w_i and μ are system variables, as in the original axioms.
4. Return the clustering that this algorithm computes as an intermediate step.

We now analyze the algorithm. We first discuss the new constraint. The universal quantifier over $v \in \mathbb{R}^d$ can be modeled by requiring that there exists a sum-of-squares proof in v of the constraint (see Fleming et al. (2019)). For the random variable $\Sigma^{-1/2}\boldsymbol{y}$, which is distributed according to a mixture of spherical Gaussians with covariance matrix I_d, it follows by standard arguments (see Hopkins and Li (2018)) that, for our choice of n, with high probability

$$\left|\frac{v}{2t}\right. \frac{1}{\alpha n}\sum_{i=1}^{n} w_i \langle \Sigma^{-1/2}(y_i-\mu), v\rangle^{2t} \leq 2(2t)^t \|v\|^{2t}$$

when α is the fraction of samples coming from one of the components, w_i is 1 for all samples from that component and 0 for all other samples, and μ is the mean of that component. Then, by a change of variables $v \to \Sigma^{1/2}v$,

$$\left|\frac{v}{2t}\right. \frac{1}{\alpha n}\sum_{i=1}^{n} w_i \langle y_i-\mu, v\rangle^{2t} \leq 2(2t)^t (v^\top \Sigma v)^t.$$

We connect now this to D. By Lemma 50, for our choice of n, with high probability

$$\left|\frac{v}{2t}\right. \frac{1}{2}\cdot \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} \leq t^t \langle v^{\otimes t}, Dv^{\otimes t}\rangle \leq 2\cdot \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}.$$

By combining this result with Lemma 49,

$$\left|\frac{v}{2t}\right. \frac{1}{2\cdot 4^t}(v^\top \Sigma v)^t \leq \langle v^{\otimes t}, Dv^{\otimes t}\rangle \leq 2\cdot 4^t \cdot (v^\top \Sigma v)^t.$$

Therefore,

$$\left|\frac{v}{2t}\right. \frac{1}{\alpha n}\sum_{i=1}^{n} w_i \langle y_i-\mu, v\rangle^{2t} \leq 4(8t)^t \langle v^{\otimes t}, Dv^{\otimes t}\rangle,$$

so the constraint is valid.

To study the guarantees of the algorithm, we show that the new constraint implies a constraint of the form required by the original algorithm, which only includes a term $\|v\|^{2t}$ on the right-hand side. We use that $\left|\frac{v}{2t}\right. \langle v^{\otimes t}, Dv^{\otimes t}\rangle \leq 2\cdot 4^t \cdot (v^\top \Sigma v)^t$ in

$$\left|\frac{v}{2t}\right. \frac{1}{\alpha n}\sum_{i=1}^{n} w_i \langle y_i-\mu, v\rangle^{2t} \leq 4(8t)^t \langle v^{\otimes t}, Dv^{\otimes t}\rangle$$

to obtain

$$\left|\frac{v}{2t}\right. \frac{1}{\alpha n}\sum_{i=1}^{n} w_i \langle y_i-\mu, v\rangle^{2t} \leq 8(32t)^t (v^\top \Sigma v)^t.$$

By a change of variables $v \to \Sigma^{-1/2}v$,

$$\left|\frac{v}{2t}\right. \frac{1}{\alpha n}\sum_{i=1}^{n} w_i \langle \Sigma^{-1/2}(y_i-\mu), v\rangle^{2t} \leq 8(32t)^t \|v\|^{2t}.$$

Finally, by dividing both sides by $4\cdot 16^t$ we obtain

$$\left|\frac{v}{2t}\right. \frac{1}{\alpha n}\sum_{i=1}^{n} w_i \left\langle \frac{1}{2^{1/t}\cdot 4}\Sigma^{-1/2}(y_i-\mu), v\right\rangle^{2t} \leq 2(2t)^t \|v\|^{2t}.$$

Then the algorithm from Theorem 5.1 of Hopkins and Li (2018) behaves as if we had samples from $\frac{1}{2^{1/t}\cdot 4}\Sigma^{-1/2}\boldsymbol{y}$, which is distributed according to a mixture of well-separated spherical Gaussians with covariance matrix $\frac{1}{4^{1/t}\cdot 16}I_d$. It is easy then to verify that we inherit the guarantees of the algorithm of Hopkins and Li (2018) and, for $t = O(R^2 + \log p_{\min}^{-1})$ large enough, we return a clustering that satisfies the statement of our theorem.

The algorithm of Hopkins and Li (2018) requires $n \geq (p_{\min}^{-1})^{O(1)} \cdot d^{O(t)} = (p_{\min}^{-1})^{O(1)} \cdot d^{O(R^2+\log p_{\min}^{-1})}$, so our choice of n is large enough to satisfy this. The time complexity is dominated by the algorithm of Hopkins and Li (2018), which has a time complexity of $n^{O(t)} = n^{O(R^2+\log p_{\min}^{-1})}$.
■

Appendix F. Auxiliary results

F.1. Sum-of-squares lemmas

We first prove a number of useful SOS facts.

Lemma 51 (Restatement of Lemma A.1 in Kothari and Steurer (2017)) *For variables $X_1, ..., X_t \in \mathbb{R}$,*

$$\vdash_{t}^{X} X_1 \cdot ... \cdot X_t \leq \frac{1}{t}(X_1^t + ... + X_t^t).$$

Lemma 52 (Restatement of Lemma A.2 in Kothari and Steurer (2017)) *For variables $A, B \in \mathbb{R}$ and $t \geq 2$ even,*

$$\vdash_{t}^{A,B} (A+B)^t \leq 2^{t-1}A^t + 2^{t-1}B^t.$$

Lemma 53 *For variables $A, B \in \mathbb{R}$ and $\delta > 0$ and $t \geq 2$ even,*

$$\vdash_{t}^{A,B} (A+B)^t \leq (1+\delta)^{t-1}A^t + \left(1+\frac{1}{\delta}\right)^{t-1} B^t.$$

Proof

$$\begin{aligned}
\vdash_{t}^{A,B} (A+B)^t &= \sum_{s=0}^{t} \binom{t}{s} A^s B^{t-s} \\
&= \sum_{s=0}^{t} \binom{t}{s} \left(\delta^{1-s/t}A\right)^s \left(\frac{1}{\delta^{s/t}}B\right)^{t-s} \\
&\overset{(1)}{\leq} \sum_{s=0}^{t} \binom{t}{s} \left(\frac{s}{t}\left(\delta^{1-s/t}A\right)^t + \frac{t-s}{s}\left(\frac{1}{\delta^{s/t}}B\right)^t\right) \\
&= \left(\sum_{s=0}^{t} \binom{t}{s} \frac{s}{t}\delta^{t-s}\right) A^t + \left(\sum_{s=0}^{t} \binom{t}{s} \frac{t-s}{t}\frac{1}{\delta^s}\right) B^t \\
&\overset{(2)}{=} (1+\delta)^{t-1}A^t + \left(1+\frac{1}{\delta}\right)^{t-1} B^t
\end{aligned}$$

where in (1) we used Lemma 51 and in (2) we used the identities

$$\sum_{s=0}^{t} \binom{t}{s} \frac{s}{t} x^{t-s} = \sum_{s=0}^{t-1} \binom{t-1}{s} x^{t-1-s} = (1+x)^{t-1}$$

and

$$\sum_{s=0}^{t} \binom{t}{s} \frac{t-s}{t} x^{s} = \sum_{s=0}^{t-1} \binom{t-1}{s} x^{s} = (1+x)^{t-1}.$$

■

Lemma 54 *For variable $X \in \mathbb{R}$ and $t \geq 0$ integer,*

$$\{X \geq 0\} \vdash_{t}^{X} \{X^t \geq 0\}.$$

Proof For t even, $\vdash_{t}^{X} X^t \geq 0$ is trivial. For t odd, we have that $\vdash_{t}^{X} X^t = X^{t-1}X \geq 0$, where we used that $\vdash_{t-1}^{X} X^{t-1} \geq 0$ because $t-1$ is even. ■

Lemma 55 *For variable $X \in \mathbb{R}$ and $t \geq 1$ integer,*

$$\{0 \leq X \leq 1\} \vdash_{t}^{X} \{X^t \leq 1\}.$$

Proof We have $\vdash_{X}^{t} 1 - X^t = (1-X)(1+X+...+X^{t-1}) \geq 0$, where we used that, by Lemma 54, $\vdash_{i}^{X} X^i \geq 0$ for $i \in \{0, ..., t-1\}$. ■

Lemma 56 (Restatement of Lemma A.3 in Kothari and Steurer (2017))
For variable $X \in \mathbb{R}$ and $t \geq 2$ even,

$$\{X^t \leq 1\} \vdash_{t}^{X} \{X \leq 1\}.$$

Lemma 57 *For variables $u, v \in \mathbb{R}^d$,*

$$\vdash_{2}^{u,v} \langle u, v\rangle^2 \leq \|u\|^2 \cdot \|v\|^2.$$

Proof By Lagrange's identity,

$$\langle u, v\rangle^2 = \|u\|^2 \cdot \|v\|^2 + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} (u_i v_j - u_j v_j)^2,$$

so

$$\vdash_{2}^{u,v} \langle u, v\rangle^2 \leq \|u\|^2 \cdot \|v\|^2.$$

■

Lemma 58 *For variable $X \in \mathbb{R}$ and $\delta \in \mathbb{R}$ and $t \geq 1$ integer,*

$$\{0 \leq X \leq 1, X^t \geq \delta\} \vdash^{X}_{t} \{X \geq \delta\}.$$

Proof We have $\vdash^{X}_{t} X = (X^t - \delta) + (1 - X)(1 + X + ... + X^{t-1}) + \delta \geq \delta$, where we used that, by Lemma 54, $\vdash^{X}_{i} X^i \geq 0$ for $i \in \{0, ..., t-1\}$. ■

Lemma 59 *For variable $X \in \mathbb{R}$ and $C \geq 2$ and $t \geq 1$ integer,*

$$\left\{0 \leq X \leq \frac{1}{Ct}\right\} \vdash^{X}_{t} \left\{(1 - X)^t \leq 1 - \frac{C-2}{C-1} tX\right\}.$$

Proof We have that

$$\begin{aligned}
\vdash^{X}_{t} (1 - X)^t &= 1 - tX + \sum_{i=2}^{t} \binom{t}{i} (-1)^i X^i \overset{(1)}{\leq} 1 - tX + \sum_{i=2}^{t} \binom{t}{i} X^i \\
&\overset{(2)}{\leq} 1 - tX + \sum_{i=2}^{t} t^i X^i = 1 - tX + tX \sum_{i=1}^{t-1} t^i X^i \\
&\overset{(3)}{\leq} 1 - tX + tX \sum_{i=1}^{t-1} \frac{1}{C^i} \overset{(4)}{\leq} 1 - tX + \frac{1}{C-1} tX \\
&= 1 - \frac{C-2}{C-1} tX.
\end{aligned}$$

We use throughout that, by Lemma 54, $\vdash^{X}_{i} X^i \geq 0$ for $i \in \{0, ..., t\}$. In (1) we used that $\vdash^{X}_{i} -X^i \leq X^i$. In (2) we used that $\binom{t}{i} \leq t^i$. In (3) we used that $\vdash^{X}_{1} 0 \leq X \leq \frac{1}{Ct}$ implies that $\vdash^{X}_{i+1} 0 \leq X^{i+1} \leq \frac{1}{(Ct)^i} X$. The upper bound is true because

$$\vdash^{X}_{i+1} \frac{1}{(Ct)^i} X - X^{i+1} = \left(\frac{1}{Ct} - X\right) \left(\sum_{j=0}^{i-1} \frac{1}{(Ct)^j} X^{i-1-j}\right) X \geq 0.$$

In (4) we used that $\sum_{i=1}^{t-1} \frac{1}{C^i} \leq \sum_{i=1}^{\infty} \frac{1}{C^i} = \frac{1}{1-\frac{1}{C}} - 1 = \frac{1}{C-1}$. ■

Lemma 60 (Restatement of Claim 1.5 in Raghavendra et al. (2018))
If $\tilde{\mathbb{E}}$ is a degree-d pseudo-expectation and if p, q are polynomials of degree at most $\frac{d}{2}$, then $\tilde{\mathbb{E}}[q(x) \cdot p(x)] \leq \frac{1}{2}\tilde{\mathbb{E}}[q(x)^2] + \frac{1}{2}\tilde{\mathbb{E}}[p(x)^2]$.

Lemma 61 *If $\tilde{\mathbb{E}}$ is a degree-d pseudo-expectation and if p is a polynomial of degree at most $\frac{d}{2}$, then $(\tilde{\mathbb{E}}[p(x)])^2 \leq \tilde{\mathbb{E}}[p(x)^2]$.*

Proof Let $\tilde{\mathbb{E}}_x$ be the given pseudo-expectation over x, and let $\tilde{\mathbb{E}}_{x'}$ be a copy of the given pseudo-expectaiton but over x' instead of x. Then we have

$$(\tilde{\mathbb{E}}_x[p(x)])^2 = (\tilde{\mathbb{E}}_x[p(x)])(\tilde{\mathbb{E}}_{x'}[p(x')]) = \tilde{\mathbb{E}}_{x,x'}[p(x)p(x')].$$

Then, by Lemma 60,

$$(\tilde{\mathbb{E}}_x[p(x)])^2 \le \frac{1}{2}\tilde{\mathbb{E}}_{x,x'}[p(x)^2] + \frac{1}{2}\tilde{\mathbb{E}}_{x,x'}[p(x')^2] = \tilde{\mathbb{E}}_x[p(x)^2].$$

■

Lemma 62 (Restatement of Lemma 4.5 in Barak and Steurer (2014))
If $\tilde{\mathbb{E}}$ is a degree-d pseudo-expectation over vectors u, v, then

$$\left(\tilde{\mathbb{E}}\,\|u+v\|_d^d\right)^{1/d} \le \left(\tilde{\mathbb{E}}\,\|u\|_d^d\right)^{1/d} + \left(\tilde{\mathbb{E}}\,\|v\|_d^d\right)^{1/d}.$$

We give now some sum-of-squares proofs that are more specific to our setting. The purpose of Lemma 63 and Lemma 64 is to aid in transforming some sum-of-squares proofs about polynomials $p(x)$ and $q(x)$ into sum-of-squares proofs about polynomials $p(x)^t$ and $q(x)^t$. Lemma 63 shows that, under some conditions, if $\{p(x) \ge 1\} \overset{x}{\vdash} \{q(x) \ge 1\}$, then also $\{p(x)^t \ge 1\} \overset{x}{\vdash} \{q(x)^t \ge 1\}$, while Lemma 64 shows that, again under some conditions, if $\{p(x) \le 1\} \overset{x}{\vdash} \{q(x) \ge 1\}$, then also $\{p(x)^t \le 1\} \overset{x}{\vdash} \{q(x)^t \ge 1\}$. These are used in Lemma 65 and Lemma 66, which implement sum-of-squares proofs with some polynomials raised to the t-th power.

Lemma 63 *Let $p, q : \mathbb{R} \to \mathbb{R}$ with $p(x) \ge 0$ for all $x \in \mathbb{R}$. Let $\gamma > 1$ be a real number and $t \ge 2$ be an even integer. Suppose that, for all $x \in \mathbb{R}$, $q(x) - 1 - \gamma(p(x) - 1) \ge 0$. Then, for all $x \in \mathbb{R}$,*

$$q(x)^t - 1 - \gamma(p(x)^t - 1) \ge 0.$$

Proof We consider two cases. First, suppose that $1 + \gamma(p(x) - 1) < 0$. This implies that $p(x) < 1 - \frac{1}{\gamma}$, which implies that $1 + \gamma(p(x)^t - 1) < 1 + \gamma(p(x) - 1) < 0$. Therefore $q(x)^t \ge 1 + \gamma(p(x)^t - 1)$ is satisfied trivially for t even.

Second, suppose that $1 + \gamma(p(x) - 1) \ge 0$. Then the given assumption implies that $q(x)^t \ge (1 + \gamma(p(x) - 1))^t$. Then

$$q(x)^t - 1 - \gamma(p(x)^t - 1) \ge (1 + \gamma(p(x) - 1))^t - 1 - \gamma(p(x)^t - 1).$$

To show that the expression on the right-hand side is non-negative, it suffices to show that

$$f(x) = (1 + \gamma(x - 1))^t - 1 - \gamma(x^t - 1)$$

is non-negative everywhere. For $\gamma > 1$, we have that $\lim_{x\to-\infty} f(x) = \infty$ and $\lim_{x\to\infty} f(x) = \infty$. Then, it suffices to show that $f(x)$ is non-negative at all its critical points. We have

$$\frac{d}{dx} f(x) = \gamma t(\gamma(x - 1) + 1)^{t-1} - \gamma t x^{t-1},$$

so

$$\frac{d}{dx}f(x) = 0 \iff \gamma(x-1)+1 = x \iff x = 1.$$

We have $f(1) = 0 \geq 0$. Therefore, $f(x) \geq 0$ for all $x \in \mathbb{R}$.

■

Lemma 64 *Let $p, q : \mathbb{R} \to \mathbb{R}$ for all $x \in \mathbb{R}$. Let $\gamma > 0$ be a real number and $t \geq 2$ be an even integer. Suppose that, for all $x \in \mathbb{R}$, $q(x) - 1 - \gamma(1 - p(x)) \geq 0$. Then, for all $x \in \mathbb{R}$,*

$$q(x)^t - 1 - \gamma(1 - p(x)^t) \geq 0.$$

Proof We consider two cases. First, suppose that $1 + \gamma(1 - p(x)) < 0$. This implies that $p(x) > 1 + \frac{1}{\gamma}$, which implies that $1 + \gamma(1 - p(x)^t) < 1 + \gamma(1 - p(x)) < 0$. Therefore $q(x)^t \geq 1 + \gamma(1 - p(x)^t)$ is satisfied trivially for t even.

Second, suppose that $1 + \gamma(1 - p(x)) \geq 0$. Then the given assumption implies that $q(x)^t \geq (1 + \gamma(1 - p(x)))^t$. Then

$$q(x)^t - 1 - \gamma(1 - p(x)^t) \geq (1 + \gamma(1 - p(x)))^t - 1 - \gamma(1 - p(x)^t).$$

To show that the expression on the right-hand side, it suffices to show that

$$f(x) = (1 + \gamma(1 - x))^t - 1 - \gamma(1 - x^t)$$

is non-negative everywhere. For $\gamma > 0$, we have that $\lim_{x \to -\infty} f(x) = \infty$ and $\lim_{x \to \infty} f(x) = \infty$. Then, it suffices to show that $f(x)$ is non-negative at all its critical points. We have

$$\frac{d}{dx}f(x) = \gamma t x^{t-1} - \gamma t(\gamma(1 - x) + 1)^{t-1},$$

so

$$\frac{d}{dx}f(x) = 0 \iff x = \gamma(1 - x) + 1 \iff x = 1.$$

We have $f(1) = 0 \geq 0$. Therefore, $f(x) \geq 0$ for all $x \in \mathbb{R}$.

■

Lemma 65, which is used in Lemma 32, provides a sum-of-squares proof of the following statement: if $\left(x^2 + \frac{1}{M}(1 - (1 - \sigma^2)x^2)\right)^t \geq \frac{1}{\gamma^t}$, then $x^{2t} \geq \left(\frac{M-\gamma}{\gamma}\frac{1}{M-1+\sigma^2}\right)^t$.

Lemma 65 *For a variable $x \in \mathbb{R}$ and for $0 \leq \sigma^2 < 1$ and $0 < \gamma < M$ and $M \geq 2$, we have that*

$$\left\{\left(\gamma\left(x^2 + \frac{1}{M}(1 - (1 - \sigma^2)x^2)\right)\right)^t \geq 1\right\} \left|\frac{x}{2t}\right. \left\{\left(\frac{\gamma}{M - \gamma}(M - 1 + \sigma^2)x^2\right)^t \geq 1\right\}.$$

Proof Let

$$p(x) = \gamma\left(x^2 + \frac{1}{M}(1 - (1 - \sigma^2)x^2)\right) = \gamma\left(\frac{M - 1 + \sigma^2}{M}x^2 + \frac{1}{M}\right)$$

and

$$q(x) = \delta(M - 1 + \sigma^2)x^2,$$

for some $\delta > 0$ to be determined later. Note that $p(x) \geq 0$ for all $x \in \mathbb{R}$.

We check now that, for all $x \in \mathbb{R}$,

$$q(x) - 1 - \frac{M\delta}{\gamma}(p(x) - 1) \geq 0,$$

which corresponds to a sum-of-squares proof that $\{p(x) \geq 1\} \big|\!\frac{x}{\;\;} \{q(x) \geq 1\}$. We note that the coefficient $\frac{M\delta}{\gamma}$ was chosen such that x^2 cancels. We have then

$$q(x) - 1 - \frac{M\delta}{\gamma}(p(x) - 1) = -1 - \frac{M\delta}{\gamma}\left(\frac{\gamma}{M} - 1\right) = \frac{M\delta}{\gamma} - 1 - \delta.$$

Set $\delta = \frac{\gamma}{M-\gamma}$, which makes the term equal to 0. Therefore, for all $x \in \mathbb{R}$,

$$q(x) - 1 - \frac{2\frac{\gamma}{M-\gamma}}{\gamma}(p(x) - 1) \geq 0.$$

Therefore, by Lemma 63, for all $x \in \mathbb{R}$,

$$f(x) = q(x)^t - 1 - \frac{2\frac{\gamma}{M-\gamma}}{\gamma}(p(x)^t - 1) \geq 0.$$

Because $f(x)$ is a univariate polynomial of degree $2t$, there also exists a sum-of-squares proof of degree at most $2t$ that $f(x) \geq 0$. Note that this constitutes a degree-$2t$ sum-of-squares proof that $\{p(x)^t \geq 1\} \big|\!\frac{x}{\;\;} \{q(x)^t \geq 1\}$. This concludes the proof.

■

Lemma 66, which is used in Lemma 33, provides a sum-of-squares proof of the following statement: if $\left(\frac{x^2+\Delta(1-(1-\sigma^2)x^2)}{1+8\Delta\sigma^2}\right)^t \leq \frac{1}{\gamma^t}$, then $x^{2t} \geq \left(\frac{\gamma\Delta-1}{\gamma(\Delta-1)}(1-10\sigma^2)\right)^t$.

Lemma 66 *For a variable $x \in \mathbb{R}$ and for $0 \leq \sigma^2 < 0.1$ and $\Delta \geq 10$ and t even and $\gamma \geq 0.9$, we have that*

$$\left\{\left(\gamma\frac{x^2 + \Delta(1 - (1 - \sigma^2)x^2)}{1 + 8\Delta\sigma^2}\right)^t \leq 1\right\} \Big|\!\frac{x}{2t} \left\{\left(\frac{\gamma(\Delta - 1)}{\gamma\Delta - 1}\frac{x^2}{1 - 10\sigma^2}\right)^t \geq 1\right\}.$$

Proof Note that we need $\sigma^2 < 0.1$ in order to have $1 - 10\sigma^2 > 0$.

Let

$$p(x) = \gamma\frac{x^2 + \Delta(1 - (1 - \sigma^2)x^2)}{1 + 8\Delta\sigma^2} = \gamma\frac{\left(1 - \Delta(1 - \sigma^2)\right)x^2 + \Delta}{1 + 8\Delta\sigma^2}$$

and

$$q(x) = \delta\frac{x^2}{1 - 10\sigma^2},$$

for some $\delta > 0$ to be determined later.

We check now that, for all $x \in \mathbb{R}$,

$$q(x) - 1 - \frac{\delta(1+8\Delta\sigma^2)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)}(1-p(x)) \geq 0,$$

which corresponds to a sum-of-squares proof that $\{p(x) \leq 1\} \vert\overset{x}{\text{—}} \{q(x) \geq 1\}$. We note that the coefficient $\frac{\delta(1+8\Delta\sigma^2)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)}$ was chosen such that x^2 cancels. We have then

$$\begin{aligned}
&q(x) - 1 - \frac{\delta(1+8\Delta\sigma^2)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)}(1-p(x)) \\
&= -1 - \frac{\delta(1+8\Delta\sigma^2)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)}\left(1 - \frac{\gamma\Delta}{1+8\Delta\sigma^2}\right) \\
&= -1 - \frac{\delta(1+8\Delta\sigma^2)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)} \frac{1+8\Delta\sigma^2-\gamma\Delta}{1+8\Delta\sigma^2} \\
&= -1 - \frac{\delta(1+8\Delta\sigma^2-\gamma\Delta)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)} \\
&= \frac{-\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2) - \delta(1+8\Delta\sigma^2-\gamma\Delta)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)} \\
&= \frac{(-10\gamma\Delta)\sigma^4 + (11\gamma\Delta - 10\gamma - 8\delta\Delta)\sigma^2 + (\gamma\delta\Delta - \gamma\Delta + \gamma - \delta)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)}.
\end{aligned}$$

Note that the denominator is positive. Set $\delta = \frac{\gamma(\Delta-1)}{\gamma\Delta-1}$. Then the numerator, viewed as a quadratic in σ^2, has roots at 0 and at $\frac{11\gamma\Delta^2-10\gamma\Delta-8\Delta^2-3\Delta+10}{10\Delta(\gamma\Delta-1)}$. Furthermore, when the second root is positive, the quadratic is also positive for all σ^2 between the two roots. Hence, in order to prove that the expression is positive for all $0 \leq \sigma^2 < 0.1$, it suffices to show that the second root is at least 0.1 in our setting. Indeed, for all $\gamma \geq 0.9$ and all $\Delta \geq 10$, we have that $\frac{11\gamma\Delta^2-10\gamma\Delta-8\Delta^2-3\Delta+10}{10\Delta(\gamma\Delta-1)} \geq 0.1$.

Therefore, for all $x \in \mathbb{R}$,

$$f(x) = q(x) - 1 - \frac{\frac{\gamma(\Delta-1)}{\gamma\Delta-1}(1+8\Delta\sigma^2)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)}(1-p(x)) \geq 0.$$

Therefore, by Lemma 64, for all $x \in \mathbb{R}$,

$$f(x) = q(x)^t - 1 - \frac{\frac{\gamma(\Delta-1)}{\gamma\Delta-1}(1+8\Delta\sigma^2)}{\gamma(\Delta(1-\sigma^2)-1)(1-10\sigma^2)}(1-p(x)^t) \geq 0.$$

Because $f(x)$ is a univariate polynomial of degree $2t$, there also exists a sum-of-squares proof of degree at most $2t$ that $f(x) \geq 0$. Note that this constitutes a degree-$2t$ sum-of-squares proof that $\{p(x)^t \leq 1\} \vert\overset{x}{\text{—}} \{q(x)^t \geq 1\}$. This concludes the proof.

■

F.2. Finite sample lemmas

Lemma 67 (Restatement of Theorem 4 in Brubaker and Vempala (2008)) *For $n \geq C\frac{kd\log^2(d/\delta)}{\epsilon^2}$, with probability $1-\delta$,*

$$\|\operatorname{cov}(\boldsymbol{y}^0)^{-1/2}(\hat{\mathbb{E}}\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0)\| \leq \epsilon$$

and

$$\|I_d - \operatorname{cov}(\boldsymbol{y}^0)^{-1/2}\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\operatorname{cov}(\boldsymbol{y}^0)^{-1/2}\| \leq \epsilon.$$

Lemma 68 *For $n \geq C\frac{kd\log^2(d/\delta)}{\epsilon^2}$, with probability $1-\delta$,*

$$\|I_d - \widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1/2}\operatorname{cov}(\boldsymbol{y}^0)\widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1/2}\| \leq \epsilon.$$

Proof By Lemma 67,

$$\|I_d - \operatorname{cov}(\boldsymbol{y}^0)^{-1/2}\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\operatorname{cov}(\boldsymbol{y}^0)^{-1/2}\| \leq 2\epsilon.$$

Then

$$\begin{aligned}(1-\epsilon)I_d &\preceq \operatorname{cov}(\boldsymbol{y}^0)^{-1/2}\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\operatorname{cov}(\boldsymbol{y}^0)^{-1/2} \preceq (1+\epsilon)I_d,\\ (1-\epsilon)\operatorname{cov}(\boldsymbol{y}^0) &\preceq \widehat{\operatorname{cov}}(\boldsymbol{y}^0) \preceq (1+\epsilon)\operatorname{cov}(\boldsymbol{y}^0),\\ \frac{1}{1+\epsilon}\widehat{\operatorname{cov}}(\boldsymbol{y}^0) &\preceq \operatorname{cov}(\boldsymbol{y}^0) \preceq \frac{1}{1-\epsilon}\widehat{\operatorname{cov}}(\boldsymbol{y}^0).\end{aligned}$$

Using that $\frac{1}{1+\epsilon} \geq 1-2\epsilon$ and $\frac{1}{1-\epsilon} \leq 1+2\epsilon$ for $\epsilon \leq 1/2$,

$$\begin{aligned}(1-2\epsilon)\widehat{\operatorname{cov}}(\boldsymbol{y}^0) &\preceq \operatorname{cov}(\boldsymbol{y}^0) \preceq (1+2\epsilon)\widehat{\operatorname{cov}}(\boldsymbol{y}^0),\\ (1-2\epsilon)I_d &\preceq \widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1/2}\operatorname{cov}(\boldsymbol{y}^0)\widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1/2} \preceq (1+2\epsilon)I_d,\\ \|I_d - \widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1/2}&\operatorname{cov}(\boldsymbol{y}^0)\widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1/2}\| \leq 2\epsilon.\end{aligned}$$

■

Lemma 69 (Restatement of Lemma 22 in Moitra and Valiant (2010)) *Let the random variable $\overline{\boldsymbol{y}} \in \mathbb{R}$ be distributed according to an isotropic mixture of k one-dimensional Gaussian distributions with minimum mixing weight $p_{\min}$. Let $\overline{y}_1, ..., \overline{y}_n \in \mathbb{R}$ be generated i.i.d. according to the distribution of $\overline{\boldsymbol{y}}$. Then, with probability $1-\delta$,*

$$\left(\frac{1}{n}\sum_{i=1}^n \overline{y}_i^t - \mathbb{E}\overline{\boldsymbol{y}}^t\right)^2 \leq \frac{1}{n\delta}p_{\min}^{-O(t)}.$$

Lemma 70 *Let the random variable $\boldsymbol{y} \in \mathbb{R}^d$ be distributed according to an isotropic mixture of k d-dimensional Gaussian distributions with minimum mixing weight $p_{\min}$. Let $y_1, ..., y_n \in \mathbb{R}^d$ be generated i.i.d. according to the distribution of $\boldsymbol{y}$. Then, with probability $1-d^t\delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n y_i^{\otimes t} - \mathbb{E}\boldsymbol{y}^{\otimes t}\right\|^2 \leq \frac{1}{n\delta}(p_{\min}^{-1}d)^{O(t)}.$$

Proof The proof is similar to the proof of Lemma 22 in Moitra and Valiant (2010).

We denote by $\boldsymbol{y}^{(j)}$ the j-th coordinate of $\boldsymbol{y}$. Let $\alpha \in \mathbb{N}^d$ satisfy $\sum_{j=1}^{d} \alpha_j = t$. Let $\boldsymbol{z}^\alpha = \prod_{j=1}^{d} (\boldsymbol{y}^{(j)})^{\alpha_j}$. By Chebyshev's inequality, with probability at least $1-\delta$,

$$\left(\frac{1}{n}\sum_{i=1}^{n} z_i^\alpha - \mathbb{E}\boldsymbol{z}^\alpha\right)^2 \leq \frac{1}{\delta}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} z_i^\alpha - \mathbb{E}\boldsymbol{z}^\alpha\right)^2\right].$$

We now bound the right-hand side. Note that $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} z_i^\alpha - \mathbb{E}\boldsymbol{z}^\alpha] = 0$. Using that for independent random variables the variance of the sum is equal to the sum of the variances,

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} z_i^\alpha - \mathbb{E}\boldsymbol{z}^\alpha\right)^2\right] = \frac{1}{n}\mathbb{E}\left[(\boldsymbol{z}^\alpha - \mathbb{E}\boldsymbol{z}^\alpha)^2\right] \leq \frac{1}{n}\mathbb{E}\left[(\boldsymbol{z}^\alpha)^2\right] \leq \frac{1}{n} p_{\min}^{-O(t)}.$$

The last inequality follows by using that $\mathbb{E}(\boldsymbol{y}^{(j)})^t \leq p_{\min}^{-O(t)}$ for all j and that, for random varaibles $\boldsymbol{x}_1, ..., \boldsymbol{x}_t \in \mathbb{R}$, $|\mathbb{E}[\boldsymbol{x}_1 \cdot ... \cdot \boldsymbol{x}_t]| \leq (\mathbb{E}\boldsymbol{x}_1^t \cdot ... \cdot \mathbb{E}\boldsymbol{x}_t^t)^{1/t}$. Then, by a union bound, with probability at least $1 - d^t\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} y_i^{\otimes t} - \mathbb{E}\boldsymbol{y}^{\otimes t}\right\|^2 \leq \frac{1}{n\delta} d^t p_{\min}^{-O(t)}.$$

■

Lemma 71 *Let the random variable $\boldsymbol{y} \in \mathbb{R}^d$ be distributed according to an istotropic mixture of k d-dimensional Gaussian distributions with minimum mixing weight $p_{\min}$. Let $y_1, ..., y_n \in \mathbb{R}^d$ be generated i.i.d. according to the distribution of $\boldsymbol{y}$. Then, for $n \geq \frac{1}{\delta}$, with probability $1 - d\delta$,*

$$\frac{1}{n}\sum_{i=1}^{n} \|y_i\|^{2t} \leq (p_{\min}^{-1} d)^{O(t)}.$$

Proof Denote by $\boldsymbol{y}^{(j)}$ the j-th coordinate of $\boldsymbol{y}$. We have

$$\frac{1}{n}\sum_{i=1}^{n} \|y_i\|^{2t} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} (y_i^{(j)})^2\right)^t \leq \frac{1}{n}\sum_{i=1}^{n} d^{t-1}\sum_{j=1}^{d} (y_i^{(j)})^{2t} = d^{t-1}\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n} (y_i^{(j)})^{2t}.$$

Note that, for each j, $\boldsymbol{y}^{(j)}$ is distributed according to an isotropic mixture of k one-dimensional Gaussian distributions. By a union bound, with probability $1 - d\delta$ the result in Lemma 69 holds for each coordinate $\boldsymbol{y}^{(j)}$. Then

$$\frac{1}{n}\sum_{i=1}^{n} \|y_i\|^{2t} \leq d^{t-1}\sum_{j=1}^{d}\left(\mathbb{E}(\boldsymbol{y}^{(j)})^{2t} + \frac{1}{\sqrt{n\delta}} p_{\min}^{-O(t)}\right).$$

We have that $\mathbb{E}(\boldsymbol{y}^{(j)})^{2t} \leq p_{\min}^{-O(t)}$. Using that $n\delta \geq 1$, we get then

$$\frac{1}{n}\sum_{i=1}^{n} \|y_i\|^{2t} \leq d^t \cdot p_{\min}^{-O(t)} \leq (p_{\min}^{-1} d)^{O(t)}.$$

■

F.3. Isotropic position transformation lemmas

The setting for the following two lemmas is that of Section D.3.

Lemma 72 (See Lemma 10 in Hsu and Kakade (2013)) *We have*

- $\hat{W}\widehat{\operatorname{cov}(\boldsymbol{y}^0)}\hat{W}^\top = I_d$,
- $\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top \succ 0$,
- $W\operatorname{cov}(\boldsymbol{y}^0)W^\top = I_d$.

Proof The results are immediate by substitution. ■

Lemma 73 (See Lemma 10 in Hsu and Kakade (2013)) *Suppose that*

$$\|I_d - \operatorname{cov}(\boldsymbol{y}^0)^{-1/2}\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\operatorname{cov}(\boldsymbol{y}^0)^{-1/2}\| \le \epsilon.$$

Then

$$\|I_d - (\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{1/2}\| \le O(\epsilon)\cdot\|\operatorname{cov}(\boldsymbol{y}^0)\|\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|.$$

Proof The given assumption implies that all eigenvalues of $\operatorname{cov}(\boldsymbol{y}^0)^{-1/2}\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\operatorname{cov}(\boldsymbol{y}^0)^{-1/2}$ lie between $1-\epsilon$ and $1+\epsilon$. Hence all eigenvalues of the inverse of this matrix lie between $\frac{1}{1+\epsilon} = 1 + O(\epsilon)$ and $\frac{1}{1-\epsilon} = 1 - O(\epsilon)$. Then

$$\|I_d - \operatorname{cov}(\boldsymbol{y}^0)^{1/2}\widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1}\operatorname{cov}(\boldsymbol{y}^0)^{1/2}\| \le O(\epsilon),$$
$$(1-O(\epsilon))\cdot\operatorname{cov}(\boldsymbol{y}^0)^{-1} \preceq \widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1} \preceq (1+O(\epsilon))\cdot\operatorname{cov}(\boldsymbol{y}^0)^{-1}.$$

Then

$$\begin{aligned}\|\hat{W}\| &= \|(\hat{U}^\top\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\hat{U})^{-1/2}\hat{U}^\top\| \le \|(\hat{U}^\top\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\hat{U})^{-1/2}\| = \|\widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1/2}\| \\ &= \|\widehat{\operatorname{cov}}(\boldsymbol{y}^0)^{-1}\|^{1/2} \le ((1+O(\epsilon))\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|)^{1/2}.\end{aligned}$$

The given assumption also implies that

$$-\epsilon\cdot\operatorname{cov}(\boldsymbol{y}^0) \preceq \widehat{\operatorname{cov}}(\boldsymbol{y}^0) - \operatorname{cov}(\boldsymbol{y}^0) \preceq \epsilon\cdot\operatorname{cov}(\boldsymbol{y}^0).$$

Hence

$$\|\widehat{\operatorname{cov}}(\boldsymbol{y}^0) - \operatorname{cov}(\boldsymbol{y}^0)\| \le \epsilon\cdot\|\operatorname{cov}(\boldsymbol{y}^0)\|.$$

Using these bounds on $\|\hat{W}\|$ and $\|\widehat{\operatorname{cov}}(\boldsymbol{y}^0) - \operatorname{cov}(\boldsymbol{y}^0)\|$, together with the fact that $\hat{W}\widehat{\operatorname{cov}}(\boldsymbol{y}^0)\hat{W}^\top = I_d$, we get that

$$\begin{aligned}\|I_d - \hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top\| &= \|\hat{W}(\widehat{\operatorname{cov}}(\boldsymbol{y}^0) - \operatorname{cov}(\boldsymbol{y}^0))\hat{W}^\top\| \\ &\le \|\hat{W}\|^2\cdot\|\widehat{\operatorname{cov}}(\boldsymbol{y}^0) - \operatorname{cov}(\boldsymbol{y}^0)\| \\ &\le \epsilon\cdot(1+O(\epsilon))\cdot\|\operatorname{cov}(\boldsymbol{y}^0)\|\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\| \\ &\le O(\epsilon)\cdot\|\operatorname{cov}(\boldsymbol{y}^0)\|\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|.\end{aligned}$$

Then all eigenvalues of $\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top$ lie between $1-\delta$ and $1+\delta$, for $\delta = O(\epsilon)\cdot\|\operatorname{cov}(\boldsymbol{y}^0)\|\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|$. Hence all eigenvalues of the square root of this matrix lie between $\sqrt{1-\delta} = 1 - O(\delta)$ and $\sqrt{1+\delta} = 1 + O(\delta)$. Then

$$\|I_d - (\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{1/2}\| \le O(\epsilon)\cdot\|\operatorname{cov}(\boldsymbol{y}^0)\|\cdot\|\operatorname{cov}(\boldsymbol{y}^0)^{-1}\|.$$

■

F.4. Miscellaneous lemmas

Lemma 74 *Let $W \in \mathbb{R}^{d\times d}$ and $\Sigma \in \mathbb{R}^{d\times d}$ with $\Sigma \succ 0$ symmetric. Suppose that $W\Sigma W^\top = I_d$. Then $W = Q\Sigma^{-1/2}$ for some orthogonal matrix $Q \in \mathbb{R}^{d\times d}$.*

Proof We have

$$W\Sigma W^\top = I_d \iff (W\Sigma^{1/2})(W\Sigma^{1/2})^\top = I_d \iff W\Sigma^{1/2} = Q \iff W = Q\Sigma^{-1/2}$$

for some orthogonal matrix Q. ■

Lemma 75 *For integers $0 \le s \le t$,*

$$\binom{2t}{2s}(2t-2s-1)!! \le \binom{t}{s}(et)^{t-s},$$

$$\binom{2t}{2s}(2t-2s-1)!! \ge \binom{t}{s}(t/2)^{t-s}.$$

Proof We use the known fact that $(2t-2s-1)!! = \frac{(2t-2s)!}{2^{t-s}(t-s)!}$. Then

$$\frac{\binom{2t}{2s}(2t-2s-1)!!}{\binom{t}{s}t^{t-s}} = \frac{\frac{(2t)!}{(2s)!(2t-2s)!}\frac{(2t-2s)!}{2^{t-s}(t-s)!}}{\frac{t!}{s!(t-s)!}t^{t-s}} = \frac{(t+1)(t+2)\cdots(2t)}{(s+1)(s+2)\cdots(2s)(2t)^{t-s}}.$$

For the upper bound, we have

$$\begin{aligned}\frac{(t+1)(t+2)\cdots(2t)}{(s+1)(s+2)\cdots(2s)(2t)^{t-s}} &= \frac{(t+1)(t+2)\cdots(t+s)}{(s+1)(s+2)\cdots(2s)}\frac{(t+s+1)(t+s+2)\cdots(2t)}{(2t)^{t-s}}\\ &\le \frac{(t+1)(t+2)\cdots(t+s)}{(s+1)(s+2)\cdots(2s)}\\ &\le \left(\frac{t}{s}\right)^s \le e^{t-s},\end{aligned}$$

where in the last inequality we used that $\left(\frac{t}{s}\right)^{\frac{s}{t-s}} = \left(1+\frac{t-s}{s}\right)^{\frac{s}{t-s}} \le e$.

For the lower bound, we have

$$\begin{aligned}\frac{(t+1)(t+2)\cdots(2t)}{(s+1)(s+2)\cdots(2s)(2t)^{t-s}} &= \frac{(t+1)(t+2)\cdots(t+s)}{(s+1)(s+2)\cdots(2s)}\frac{(t+s+1)(t+s+2)\cdots(2t)}{(2t)^{t-s}}\\ &\ge \frac{(t+s+1)(t+s+2)\cdots(2t)}{(2t)^{t-s}}\\ &\ge \frac{1}{2^{t-s}}.\end{aligned}$$

■

F.5. Proofs deferred from Section B

Proof of Lemma 19 For the first proof, with probability $1 - \epsilon/100$,

$$\begin{aligned}
\Big|\frac{v}{2}\; v^\top \operatorname{cov}(\boldsymbol{z}) v &= v^\top \widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2} \widehat{\operatorname{cov}}(\boldsymbol{z})^{-1/2} \operatorname{cov}(\boldsymbol{z}) \widehat{\operatorname{cov}}(\boldsymbol{z})^{-1/2} \widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2} v \\
&\overset{(*)}{\leq} \|\widehat{\operatorname{cov}}(\boldsymbol{z})^{-1/2} \operatorname{cov}(\boldsymbol{z}) \widehat{\operatorname{cov}}(\boldsymbol{z})^{-1/2}\| \cdot \|\widehat{\operatorname{cov}}(\boldsymbol{z})^{1/2} v\|^2 \\
&\leq (1+\eta) C,
\end{aligned}$$

where in (*) we used Lemma 68.

Similarly, for the second proof, with probability $1 - \epsilon/100$

$$\begin{aligned}
\Big|\frac{v}{2}\; v^\top \widehat{\operatorname{cov}}(\boldsymbol{z}) v &= v^\top \operatorname{cov}(\boldsymbol{z})^{1/2} \operatorname{cov}(\boldsymbol{z})^{-1/2} \widehat{\operatorname{cov}}(\boldsymbol{z}) \operatorname{cov}(\boldsymbol{z})^{-1/2} \operatorname{cov}(\boldsymbol{z})^{1/2} v \\
&\overset{(*)}{\leq} \| \operatorname{cov}(\boldsymbol{z})^{-1/2} \widehat{\operatorname{cov}}(\boldsymbol{z}) \operatorname{cov}(\boldsymbol{z})^{-1/2}\| \cdot \| \operatorname{cov}(\boldsymbol{z})^{1/2} v\|^2 \\
&\leq (1+\eta) C.
\end{aligned}$$

where in (*) we used Lemma 67. ■

Proof of Lemma 20 We have

$$\Big|\frac{v}{2t}\; \hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} = \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} + \left(\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t}\right).$$

For the second term we have that

$$\begin{aligned}
&\Big|\frac{v}{4t}\; \left(\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t}\right)^2 \\
&\quad= \left(\hat{\mathbb{E}}\langle \operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z}, \operatorname{cov}(\boldsymbol{z})^{1/2} v\rangle^{2t} - \mathbb{E}\langle \operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z}, \operatorname{cov}(\boldsymbol{z})^{1/2} v\rangle^{2t}\right)^2 \\
&\quad= \left(\hat{\mathbb{E}}\langle (\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t}, (\operatorname{cov}(\boldsymbol{z})^{1/2} v)^{\otimes 2t}\rangle - \mathbb{E}\langle (\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t}, (\operatorname{cov}(\boldsymbol{z})^{1/2} v)^{\otimes 2t}\rangle\right)^2 \\
&\quad= \langle \hat{\mathbb{E}}(\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t} - \mathbb{E}(\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t}, (\operatorname{cov}(\boldsymbol{z})^{1/2} v)^{\otimes 2t}\rangle^2 \\
&\quad\leq \|\hat{\mathbb{E}}(\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t} - \mathbb{E}(\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t}\|^2 \cdot \|(\operatorname{cov}(\boldsymbol{z})^{1/2} v)^{\otimes 2t}\| \\
&\quad\leq \|\hat{\mathbb{E}}(\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t} - \mathbb{E}(\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t}\|^2 \cdot C^t
\end{aligned}$$

Note that $\mathbb{E}\boldsymbol{z} = 0$. Then $\operatorname{cov}(\boldsymbol{z})^{-1/2}\boldsymbol{z}$ is in isotropic position. By Lemma 70, with probability $1 - d^{2t}\delta$, we have that

$$\|\hat{\mathbb{E}}(\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t} - \mathbb{E}(\operatorname{cov}(\boldsymbol{z})^{-1/2} \boldsymbol{z})^{\otimes 2t}\|^2 \leq \frac{1}{n\delta}(p_{\min}^{-2} d)^{O(t)}.$$

Select $n = (C p_{\min}^{-1} d)^{O(t)} \eta^{-1} \epsilon^{-1}$ large enough the right-hand side is upper boundeed by $\eta^2 C^{-t}$ with probability at least $1 - \epsilon$. Then it follows that

$$\Big|\frac{v}{4t}\; \left(\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t}\right)^2 \leq \eta^2.$$

Then, by Lemma 56, we get that

$$\Big|\frac{v}{O(t)}\; \hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq \eta,$$

$$\left|\frac{v}{O(t)}\right. -\hat{\mathbb{E}}\langle \boldsymbol{z}, v\rangle^{2t} + \mathbb{E}\langle \boldsymbol{z}, v\rangle^{2t} \leq \eta.$$

Rearranging leads to the desired results.

■

F.6. Proofs deferred from Section C

Proof of Lemma 35 We have

$$\left|\frac{v}{2t}\right. \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} = \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} + \left(\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}\right).$$

For the second term we have that

$$\begin{aligned}\left|\frac{v}{4t}\right. \left(\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}\right)^2 &= \left(\hat{\mathbb{E}}\langle \boldsymbol{y}^{\otimes 2t}, v^{\otimes 2t}\rangle - \mathbb{E}\langle \boldsymbol{y}^{\otimes 2t}, v^{\otimes 2t}\rangle\right)^2 \\ &= \langle \hat{\mathbb{E}}\boldsymbol{y}^{\otimes 2t} - \mathbb{E}\boldsymbol{y}^{\otimes 2t}, v^{\otimes 2t}\rangle^2 \\ &\leq \|\hat{\mathbb{E}}\boldsymbol{y}^{\otimes 2t} - \mathbb{E}\boldsymbol{y}^{\otimes 2t}\|^2.\end{aligned}$$

By Lemma 70, with probability $1 - d^{2t}\delta$, we have that

$$\|\hat{\mathbb{E}}\boldsymbol{y}^{\otimes 2t} - \mathbb{E}\boldsymbol{y}^{\otimes 2t}\|^2 \leq \frac{1}{n\delta}(p_{\min}^{-1}d)^{O(t)}.$$

For $n \geq (p_{\min}^{-1}d)^{O(t)}\eta^{-2}\epsilon^{-1}$ the right-hand side is η^2 with probability at least $1-\epsilon$. Then it follows that

$$\left|\frac{v}{4t}\right. \left(\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}\right)^2 \leq \eta^2.$$

Then, by Lemma 56, we get that

$$\left|\frac{v}{O(t)}\right. \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} \leq \eta,$$

$$\left|\frac{v}{O(t)}\right. -\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} + \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} \leq \eta.$$

Rearranging leads to the desired results. ■

F.7. Proofs deferred from Section D

Proof of Lemma 43 We have

$$\left|\frac{v}{2t}\right. \hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t} = \hat{\mathbb{E}}\left(\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle + \langle \hat{W}(\mathbb{E}\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle\right)^{2t}.$$

Let $\delta > 0$ to be specified later. For the upper bound:

$$\begin{aligned}&\left|\frac{v}{2t}\right. \hat{\mathbb{E}}\left(\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle + \langle \hat{W}(\mathbb{E}\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle\right)^{2t} \\ &\quad\leq (1+\delta)^{2t-1} \cdot \hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0 - \mathbb{E}\boldsymbol{y}^0), v\rangle^{2t} + \left(1 + \frac{1}{\delta}\right)^{2t-1} \cdot \langle \hat{W}(\mathbb{E}\boldsymbol{y}^0 - \hat{\mathbb{E}}\boldsymbol{y}^0), v\rangle^{2t},\end{aligned}$$

where in the inequality we used Lemma 53.

For the lower bound:

$$\begin{aligned}
&\left|\frac{v}{2t}\right. \hat{\mathbb{E}}\left(\langle \hat{W}(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\rangle+\langle \hat{W}(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0),v\rangle\right)^{2t}\\
&\overset{(1)}{\geq}\left(\frac{1}{1+\delta}\right)^{2t-1}\cdot\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\rangle^{2t}-\left(\frac{1+\frac{1}{\delta}}{1+\delta}\right)^{2t-1}\cdot\langle \hat{W}(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0),v\rangle^{2t}\\
&\overset{(2)}{\geq}\left(\frac{1}{1+\delta}\right)^{2t-1}\cdot\hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\rangle^{2t}-O\left(1+\frac{1}{\delta}\right)^{2t-1}\cdot\langle \hat{W}(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0),v\rangle^{2t},
\end{aligned}$$

where in (1) we used that, by Lemma 53, $\left|\frac{A,B}{2t}\right. A^{2t}\leq(1+\delta)^{2t-1}(A+B)^{2t}+(1+\frac{1}{\delta})^{2t-1}B^{2t}$, so $\left|\frac{A,B}{2t}\right. (A+B)^{2t}\geq(\frac{1}{1+\delta})^{2t-1}A^{2t}-(\frac{1+\frac{1}{\delta}}{1+\delta})^{2t-1}B^{2t}$. In (2) we assumed that $\delta=O(1)$, which will be the case for our choice.

Now take $\delta=\frac{\eta}{100t}$. Then $(1+\delta)^{2t-1}\leq1+\eta$ and $\left(\frac{1}{1+\delta}\right)^{2t-1}\geq1-\eta$ for η small.

For the second term in both bounds, we use that

$$\begin{aligned}
\left|\frac{v}{2t}\right. \langle \hat{W}(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0),v\rangle^{2t}&\leq\|\hat{W}(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0)\|^{2t}\\
&=\|\hat{W}W^{-1}W(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0)\|^{2t}\\
&\leq\|\hat{W}W^{-1}\|^{2t}\cdot\|W(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0)\|^{2t}\\
&=\|(\hat{W}\operatorname{cov}(\boldsymbol{y}^0)^{-1}\hat{W}^\top)^{1/2}\|^{2t}\cdot\|W(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0)\|^{2t}.
\end{aligned}$$

By Lemma 67 and Lemma 73, with probability $1-\epsilon$,

$$\|W(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0)\|\leq\left(\frac{\eta}{t}\right)^{O(1)}$$

and

$$\|(\hat{W}\operatorname{cov}(\boldsymbol{y}^0)^{-1}\hat{W}^\top)^{1/2}\|\leq1.$$

Then the second term in both bounds becomes

$$\left|\frac{v}{2t}\right. O\left(1+\frac{1}{\delta}\right)^{2t-1}\cdot\langle \hat{W}(\mathbb{E}\boldsymbol{y}^0-\hat{\mathbb{E}}\boldsymbol{y}^0),v\rangle^{2t}\leq O\left(\frac{100t}{\eta}\right)^{2t-1}\cdot\left(\frac{\eta}{t}\right)^{O(t)}\leq\eta.$$

■

Proof of Lemma 44 We have

$$\left|\frac{v}{2t}\right. \hat{\mathbb{E}}\langle \hat{W}(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\rangle^{2t}=\hat{\mathbb{E}}\left(\langle W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\rangle+\left\langle\left(\hat{W}-W\right)(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle\right)^{2t}.$$

Let $\delta>0$ to be specified later. For the upper bound:

$$\begin{aligned}
&\left|\frac{v}{2t}\right. \hat{\mathbb{E}}\left(\langle W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\rangle+\left\langle\left(\hat{W}-W\right)(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle\right)^{2t}\\
&\quad\leq(1+\delta)^{2t-1}\cdot\hat{\mathbb{E}}\left\langle W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t}+\left(1+\frac{1}{\delta}\right)^{2t-1}\cdot\hat{\mathbb{E}}\left\langle\left(\hat{W}-W\right)(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t},
\end{aligned}$$

where in the inequality we used Lemma 53.

For the lower bound:

$$\begin{aligned}
&\left|\frac{v}{2t}\right. \hat{\mathbb{E}}\left(\left\langle W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle+\left\langle\left(\hat{W}-W\right)(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle\right)^{2t}\\
&\overset{(1)}{\geq}\left(\frac{1}{1+\delta}\right)^{2t-1}\cdot\hat{\mathbb{E}}\left\langle W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t}-\left(\frac{1+\frac{1}{\delta}}{1+\delta}\right)^{2t-1}\cdot\hat{\mathbb{E}}\left\langle\left(\hat{W}-W\right)(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t}\\
&\overset{(2)}{\geq}\left(\frac{1}{1+\delta}\right)^{2t-1}\cdot\hat{\mathbb{E}}\left\langle W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t}-O\left(1+\frac{1}{\delta}\right)^{2t-1}\cdot\hat{\mathbb{E}}\left\langle\left(\hat{W}-W\right)(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t},
\end{aligned}$$

where in (1) we used that, by Lemma 53, $\left|\frac{A,B}{2t}\right. A^{2t}\leq(1+\delta)^{2t-1}(A+B)^{2t}+(1+\frac{1}{\delta})^{2t-1}B^{2t}$, so $\left|\frac{A,B}{2t}\right. (A+B)^{2t}\geq(\frac{1}{1+\delta})^{2t-1}A^{2t}-(\frac{1+\frac{1}{\delta}}{1+\delta})^{2t-1}B^{2t}$. In (2) we assumed that $\delta=O(1)$, which will be the case for our choice.

Now take $\delta=\frac{\eta}{100t}$. Then $(1+\delta)^{2t-1}\leq 1+\eta$ and $\left(\frac{1}{1+\delta}\right)^{2t-1}\geq 1-\eta$ for η small.

For the second term in both bounds, we use that

$$\begin{aligned}
\left|\frac{v}{2t}\right. \hat{\mathbb{E}}\left\langle\left(\hat{W}-W\right)(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t}&=\hat{\mathbb{E}}\left\langle\left(\hat{W}W^{-1}-I_d\right)W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t}\\
&\leq\left\|I_d-\hat{W}W^{-1}\right\|^{2t}\cdot\hat{\mathbb{E}}\left\|W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0)\right\|^{2t}\\
&=\left\|I_d-(\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{1/2}\right\|^{2t}\cdot\hat{\mathbb{E}}\left\|W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0)\right\|^{2t}.
\end{aligned}$$

By Lemma 67 and Lemma 73, with probability $1-\epsilon$,

$$\left\|I_d-(\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{1/2}\right\|\leq\left(\frac{\eta}{tp_{\min}^{-1}d}\right)^{O(1)}.$$

By Lemma 71, with probability $1-\epsilon$,

$$\hat{\mathbb{E}}\left\|W(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0)\right\|^{2t}\leq(p_{\min}^{-1}d)^{O(t)}.$$

Then the second term in both bounds becomes

$$\begin{aligned}
&\left|\frac{v}{2t}\right. O\left(1+\frac{1}{\delta}\right)^{2t-1}\cdot\hat{\mathbb{E}}\left\langle\left(\hat{W}-W\right)(\boldsymbol{y}^0-\mathbb{E}\boldsymbol{y}^0),v\right\rangle^{2t}\\
&\qquad\leq O\left(\frac{100t}{\eta}\right)^{2t-1}\cdot\left(\frac{\eta}{tp_{\min}^{-1}d}\right)^{O(t)}\cdot(p_{\min}^{-1}d)^{O(t)}\leq\eta.
\end{aligned}$$

■

Proof of Lemma 45 We have

$$\begin{aligned}
|\langle W(\mu_i^0-\mu_j^0),v\rangle-\langle\hat{W}(\mu_i^0-\mu_j^0),v\rangle|&=|\langle(W-\hat{W})(\mu_i^0-\mu_j^0),v\rangle|\\
&=|\langle(I_d-\hat{W}W^{-1})W(\mu_i^0-\mu_j^0),v\rangle|\\
&\leq\|I_d-\hat{W}W^{-1}\|\cdot\|W(\mu_i^0-\mu_j^0)\|\\
&=\|I_d-(\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{1/2}\|\cdot\|W(\mu_i^0-\mu_j^0)\|.
\end{aligned}$$

By Lemma 67 and Lemma 73, with probability $1-\epsilon$,

$$\left\|I_d - (\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{1/2}\right\| \le \left(\frac{\eta}{p_{\min}^{-1}}\right)^{O(1)}.$$

Note that $W(\mu_i^0 - \mu_j^0) = \mu_i - \mu_j = \langle \mu_i - \mu_j, u\rangle u$, so $\|W(\mu_i^0 - \mu_j^0)\| = |\langle \mu_i - \mu_j, u\rangle|$. Using that $\sum_{i=1}^k p_i \langle \mu_i, u\rangle^2 \le 1$, we have that $\sum_{i=1}^k \langle \mu_i, u\rangle^2 \le p_{\min}^{-1}$, so $\langle \mu_i, u\rangle^2 \le p_{\min}^{-1}$, so $|\langle \mu_i - \mu_j, u\rangle| \le 2\sqrt{p_{\min}^{-1}}$. Then $\|W(\mu_i^0 - \mu_j^0)\| \le 2\sqrt{p_{\min}^{-1}}$.

Therefore,

$$|\langle W(\mu_i^0 - \mu_j^0), v\rangle - \langle \hat{W}(\mu_i^0 - \mu_j^0), v\rangle| \le \left(\frac{\eta}{p_{\min}^{-1}}\right)^{O(1)} \cdot 2\sqrt{p_{\min}^{-1}} \le \eta.$$

■

Proof of Lemma 45 [Proof of Lemma 46.] We have

$$\begin{aligned}
\|v^\top W(\Sigma^0)^{1/2} - v^\top \hat{W}(\Sigma^0)^{1/2}\| &= \|v^\top (W - \hat{W})(\Sigma^0)^{1/2}\| \\
&= \|v^\top (I_d - \hat{W}W^{-1})W(\Sigma^0)^{1/2}\| \\
&\le \|I_d - \hat{W}W^{-1}\| \cdot \|W(\Sigma^0)^{1/2}\| \\
&= \|I_d - (\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{1/2}\| \cdot \|W(\Sigma^0)^{1/2}\|.
\end{aligned}$$

By Lemma 67 and Lemma 73, with probability $1-\epsilon$,

$$\left\|I_d - (\hat{W}\operatorname{cov}(\boldsymbol{y}^0)\hat{W}^\top)^{1/2}\right\| \le \eta.$$

Note that, by Lemma 74, $W(\Sigma^0)^{1/2} = Q(\Sigma^0)^{-1/2}(\Sigma^0)^{1/2} = Q$ for an orthogonal matrix Q. We have $\|Q\| = 1$, so $\|W(\Sigma^0)^{1/2}\| = 1$.

Therefore,

$$\|v^\top W(\Sigma^0)^{1/2} - v^\top \hat{W}(\Sigma^0)^{1/2}\| \le \eta.$$

■

F.8. Proofs deferred from Section E

Proof of Lemma 50 We have

$$\begin{aligned}
\Big|\frac{v}{2t}\,\hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} &= \hat{\mathbb{E}}\langle \boldsymbol{y}^{\otimes 2t}, v^{\otimes 2t}\rangle - \mathbb{E}\langle \boldsymbol{y}^{\otimes 2t}, v^{\otimes 2t}\rangle \\
&= \langle \hat{\mathbb{E}}\boldsymbol{y}^{\otimes 2t} - \mathbb{E}\boldsymbol{y}^{\otimes 2t}, v^{\otimes 2t}\rangle \\
&= \langle \hat{\mathbb{E}}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} - \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}, (vv^\top)^{\otimes t}\rangle.
\end{aligned}$$

We now bound $\hat{\mathbb{E}}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} - \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}$. Define

$$E = \hat{\mathbb{E}}(\operatorname{cov}(\boldsymbol{y})^{-1/2}\boldsymbol{y}\boldsymbol{y}^\top \operatorname{cov}(\boldsymbol{y})^{-1/2})^{\otimes t} - \mathbb{E}(\operatorname{cov}(\boldsymbol{y})^{-1/2}\boldsymbol{y}\boldsymbol{y}^\top \operatorname{cov}(\boldsymbol{y})^{-1/2})^{\otimes t}.$$

By Lemma 70, with probability $1 - d^{2t}\delta$, we have that

$$\|E\|_F = \|\hat{\mathbb{E}}(\mathrm{cov}(\boldsymbol{y})^{-1/2}\boldsymbol{y})^{\otimes 2t} - \mathbb{E}(\mathrm{cov}(\boldsymbol{y})^{-1/2}\boldsymbol{y})^{\otimes 2t}\| \leq \frac{1}{\sqrt{n\delta}}(p_{\min}^{-1}d)^{O(t)}.$$

For $n \geq (p_{\min}^{-1}d)^{O(t)}\eta^{-2}\epsilon^{-1}$ this term is η with probability at least $1-\epsilon$. In this case $\|E\| \leq \|E\|_F \leq \eta$, so

$$-\eta \cdot \mathrm{cov}(\boldsymbol{y})^{\otimes t} \preceq (\mathrm{cov}(\boldsymbol{y})^{1/2})^{\otimes t} E (\mathrm{cov}(\boldsymbol{y})^{1/2})^{\otimes t} \preceq \eta \cdot \mathrm{cov}(\boldsymbol{y})^{\otimes t}.$$

We observe the connection between $\hat{\mathbb{E}}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} - \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}$ and E:

$$\hat{\mathbb{E}}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} - \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} = (\mathrm{cov}(\boldsymbol{y})^{1/2})^{\otimes t} E (\mathrm{cov}(\boldsymbol{y})^{1/2})^{\otimes t}.$$

Using this and using that $\mathrm{cov}(\boldsymbol{y}) \preceq \mathbb{E}\boldsymbol{y}\boldsymbol{y}^\top$, we finally obtain that

$$-\eta \cdot \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} \preceq \hat{\mathbb{E}}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} - \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} \preceq \eta \cdot \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}.$$

Then

$$\begin{aligned}\left|\tfrac{v}{2t}\right. \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} &= \langle \hat{\mathbb{E}}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} - \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}, (vv^\top)^{\otimes t}\rangle \\ &\leq \eta \cdot \langle \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}, (vv^\top)^{\otimes t}\rangle \\ &= \eta \cdot \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}\end{aligned}$$

and

$$\begin{aligned}\left|\tfrac{v}{2t}\right. \hat{\mathbb{E}}\langle \boldsymbol{y}, v\rangle^{2t} - \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t} &= \langle \hat{\mathbb{E}}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t} - \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}, (vv^\top)^{\otimes t}\rangle \\ &\geq -\eta \cdot \langle \mathbb{E}(\boldsymbol{y}\boldsymbol{y}^\top)^{\otimes t}, (vv^\top)^{\otimes t}\rangle \\ &= -\eta \cdot \mathbb{E}\langle \boldsymbol{y}, v\rangle^{2t}.\end{aligned}$$

The conclusion follows. ■