Learning GMMs with Nearly Optimal Robustness Guarantees

Allen Liu CLIU568@MIT.EDU
Massachusetts Institute of Technology

Ankur Moitra MOITRA@MIT.EDU
Massachusetts Institute of Technology

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

In this work we solve the problem of robustly learning a high-dimensional Gaussian mixture model with k components from ϵ-corrupted samples up to accuracy $\widetilde{O}(\epsilon)$ in total variation distance for any constant k and with mild assumptions on the mixture. This robustness guarantee is optimal up to polylogarithmic factors. The main challenge is that most earlier works rely on learning individual components in the mixture, but this is impossible in our setting, at least for the types of strong robustness guarantees we are aiming for. Instead we introduce a new framework which we call *strong observability* that gives us a route to circumvent this obstacle.

Keywords: GMMs, Mixture of Gaussians, Density Estimation, Robust Statistics

1. Introduction

Gaussian mixture models have a long and storied history. They were first introduced in the pioneering work of Pearson (1894) and have found wide-ranging applications ever since, as a natural model for data believed to be coming from two or more heterogeneous sources. Early works focused on the statistical complexity (Teicher, 1961), namely bounding the number of samples needed to estimate the Gaussian mixture model to within some desired accuracy. More recently, these problems have been revisited with an emphasis on giving computationally efficient algorithms that work in high dimensions and with minimal assumptions (Dasgupta, 1999; Kalai et al., 2010; Moitra and Valiant, 2010; Belkin and Sinha, 2010; Ge et al., 2015).

There are different types of learning goals we could ask for:

(1) In *parameter learning*, we want to estimate the mixture on a component-by-component basis. We ask that there is a matching between the components in our hypothesis and those of the true mixture so that across the matching we are close in total variation distance and get the mixing weights approximately correct. Alternatively we could ask to be close in an appropriate parameter distance instead.

(2) In *proper density estimation*, we relax the goal of estimating the individual components. Rather, we want to output a hypothesis from the correct family (i.e. a Gaussian mixture model) and we require that it is statistically close as a distribution to the true mixture.

(3) Finally in *improper density estimation* the setup is the same as above except that we allow ourselves to output any hypothesis, even if it is not from the family we are trying to learn.

The distinctions between these notions of learning will play a key role in this work.

Recently, researchers have begun to revisit many of the key problems in high-dimensional learning from the perspective of robust statistics (Diakonikolas et al., 2019a, 2017a; Lai et al., 2016; Charikar et al., 2017; Balakrishnan et al., 2017; Klivans et al., 2018; Diakonikolas et al., 2019b; Hopkins and Li, 2018; Kothari et al., 2018; Li, 2018; Steinhardt, 2018; Diakonikolas and Kane, 2019; Bakshi and Prasad, 2021; Chen et al., 2020). In particular, we allow an adversary to arbitrarily corrupt an ϵ-fraction of the samples. In this setting, it is no longer possible to learn the original distribution to any desired accuracy. In fact the algorithmic problems associated with working in high-dimensions become even more acute in the sense that many algorithms that work in the non-robust setting turn out to only be able to tolerate a fraction of corruptions that decays inverse polynomially with the dimension.

For the most part, the emphasis in algorithmic robust statistics has been on getting some dimension-independent robustness guarantee. And only for simpler problems, like estimating a high-dimensional Gaussian (Diakonikolas et al., 2019a, 2018) and linear regression (Bakshi and Prasad, 2021) are nearly optimally robust algorithms known. In this work, we will take aim at the problem of giving algorithms with nearly optimal robustness guarantees for the challenging task of learning Gaussian mixture models. Most relevant to us are the recent works of Liu and Moitra (2021) and Bakshi et al. (2020) who gave the first robust algorithms for learning Gaussian mixture models that achieve dimension-independent robustness guarantees. Let k be the number of components in the mixture. These works achieve error rates of $\epsilon^{\Omega_k(1)}$. However in terms of the quantitative dependence on ϵ, these works are far from optimal, and here we will ask for much more:

> *Given ϵ-corrupted samples, is there an efficient algorithm for estimating the true mixture within $\widetilde{O}(\epsilon)$ in total variation distance for any constant k?*

Such a bound would be optimal up to polylogarithmic factors. And even in the case of a single Gaussian it is known that there are fundamental tensions between robustness and computational efficiency, and there is evidence that it might not be possible to obtain $O(\epsilon)$ accuracy (at least in a subtractive model of noise) (Diakonikolas et al., 2017b). As we will discuss below, in order to solve this problem we will need new frameworks and strategies that avoid trying to learn individual components.

1.1. Observability

Our main conceptual contribution is a new framework, which we call *observability*, for framing high-dimensional density estimation problems. We use this notion as a building block for how to design algorithms for robustly learning a mixture of Gaussians even when it is impossible to learn it on a component-by-component basis. Observability involves having a set of test functions $f_1, \dots, f_n$ that are used to measure a distribution in a family $\mathcal{F}$. It has many parallels, but also important differences, with the commonly used notion of identifiability. We begin with a definition.

Definition 1 (Observability) *Given a family of distributions $\mathcal{F}$ and a set of test functions $f_1, \dots, f_n$, we say that the family $\mathcal{F}$ is observable through the test functions $f_1, \dots, f_n$ if any two distributions $\mathcal{M}, \mathcal{M}' \in \mathcal{F}$ that produce identical measurements (i.e. $\mathbb{E}_{\mathcal{M}}[f_i] = \mathbb{E}_{\mathcal{M}'}[f_i]$), they must be equivalent as distributions.*

In other words, $\mathcal{F}$ is observable through a family of test functions $f_1, \dots, f_n$ if these test function measurements uniquely determine a distribution in $\mathcal{F}$. Of course we need a more algorithmically useful version of observability with quantitative guarantees. Since our goal is to achieve nearly optimal robustness guarantees, we will need the test function discrepancy and the TV distance to be proxies for each other up to logarithmic factors. We call this *strong observability*.

Definition 2 (Strong Observability) *Given a family of distributions $\mathcal{F}$ and a set of test functions $f_1, \dots, f_n$, we say that the family $\mathcal{F}$ is strongly observable through the test functions $f_1, \dots, f_n$ if for any two distributions $\mathcal{M}, \mathcal{M}' \in \mathcal{F}$, we have*

$$d_{TV}(\mathcal{M}, \mathcal{M}') \cong \sum_i |\mathbb{E}_{\mathcal{M}}[f_i] - \mathbb{E}_{\mathcal{M}'}[f_i]|$$

where the $\cong$ means that the two sides are equivalent up to logarithmic factors.

Strong observability is a subtle property. There are related facts that are much easier to establish: For any two mixtures $\mathcal{M}$ and $\mathcal{M}'$ that are ϵ-far in total variation distance, there is a test function f (whose values are bounded between zero and one) where $|\mathbb{E}_{\mathcal{M}}[f] - \mathbb{E}_{\mathcal{M}'}[f]| \geq \epsilon$. However the quantifiers are in the wrong order. In particular, the particular test function that distinguishes $\mathcal{M}$ and $\mathcal{M}'$ could vary arbitrarily and pathologically as we vary the two mixtures. In contrast, what makes our notion of observability algorithmically useful is that the family of test functions is fixed in advance and we will show that a polynomial number of them suffice. Thus, for a density estimation algorithm, our problem is reduced to measuring the test functions on the true mixture and then computing any distribution in $\mathcal{F}$ that matches these measurements. Strong observability implies that this strategy will obtain nearly optimal robustness guarantees, even in the face of adversarial corruptions. The notion of observability seems natural and fundamental to high-dimensional learning, but as far as we are aware it has not appeared in the literature before.

Our main result is in establishing that strong observability is possible for GMMs with a polynomial number of test functions, provided that the components are in regular form (see Definition 24). Roughly, a mixture is in regular form if all components are not too poorly conditioned and not too separated from each other. Moreover we can reduce the general learning problem to the regular form case by invoking recent results on robust clustering. Our result can be summarized as:

> *A mixture of a constant number of Gaussians in regular form is strongly observable through constant degree Hermite moments* [1].

More specifically we prove:

Theorem 3 (Informal, see Theorem 50) *For two mixtures $\mathcal{M}, \mathcal{M}'$ of a constant number of Gaussians in regular form, the distance between their first $O_k(1)$ Hermite moments and their TV distance are equivalent up to logarithmic factors.*

While our overall algorithm builds on the line of previous works on robustly learning GMMs (Diakonikolas et al., 2020a; Bakshi and Kothari, 2020; Liu and Moitra, 2021), our key contribution is the proof of strong observability. It leverages the recent generating function technology in Liu and Moitra (2021) but in new ways that avoid using the sum-of-squares hierarchy.

[1] Hermite moments will be defined more formally in the next section, but for now can be thought of modifications of standard moments that can be robustly estimated.

In the appendix (Section A), we provide a further discussion on observability and in particular a comparison with the more familiar concept of identifiability, which is usually thought of as the crucial ingredient in parameter learning algorithms. As we discuss, it is *information theoretically impossible* to learn the parameters to accuracy better than $\epsilon^{O(1/k)}$. Thus, approaches based on comparing parameters or identifiability appear doomed at this barrier and cannot obtain the types of strong robustness guarantees that we are hoping for.

1.2. Our Results and Techniques

Our main result is an algorithm whose robustness guarantees are optimal up to logarithmic factors for any constant k. The formal statement can be found in Theorem 72.

Theorem 4 (Informal) *Let k be a constant. Let $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ be a mixture of Gaussians in $\mathbb{R}^d$ whose components have variances lower and upper bounded in all directions and such that the mixing weights are lower bounded (both of these bounds can be any function of k). Given $\mathrm{poly}(d/\epsilon)$ samples from $\mathcal{M}$ that are ϵ-corrupted, there is an algorithm that runs in time $\mathrm{poly}(n)$ and with high probability outputs a distribution f that is a mixture of k polynomial Gaussians (see below for formal definition) such that*

$$d_{\mathsf{TV}}(\mathcal{M}, f) \leq \widetilde{O}(\epsilon)\,.$$

Remark 5 *While weaker than the familiar goal of parameter learning (which as we discussed is **impossible** for the quantitative robustness guarantees we are aiming for), improper density estimation still has important applications. For example, our estimate f will have the property that we can efficiently sample from it, which means we can compute probabilities of arbitrary events and various statistics without needing new samples from $\mathcal{M}$.*

Remark 6 *Our result does require slightly stronger assumptions on the mixture than the previous works Liu and Moitra (2021); Bakshi et al. (2020). The first assumption about the components having bounded variances in all directions is a technical nuance that arises because if there is some component with identity covariance and some other component with variance $\epsilon^{0.1}$ in some direction and identity covariance in all other directions, then we cannot "separate" these components via clustering as then we will lose a $\epsilon^{\Omega(1)}$ factor in our accuracy guarantee. On the other hand, if we don't separate such components, then we will lose $\epsilon^{\Omega(1)}$ factors in our quantitative relationship between Hermite moments and TV distance (strong observability). This issue does not arise in previous works because they only aim for $\epsilon^{\Omega(1)}$ accuracy in which case we can simply separate such components via clustering. The second assumption about mixing weights is inherited from Liu and Moitra (2021) because we need to first obtain rough estimates for all of the components and we cannot do this if there are small mixing weights. These two assumptions are not information-theoretically necessary and we leave it as an open question to see if it is possible to remove them.*

As we discussed previously, the main challenge is that in our setting, parameter learning is not information-theoretically possible, at least not with the kinds of robustness guarantees that we are aiming for. Our algorithm and its analysis revolve around showing that a Gaussian mixture model is strongly observable through its Hermite moments. We now sketch the proof.

First, we give some background. For a distribution D, the characteristic function is defined as $\widehat{D}(X) = \mathbb{E}_{z \sim D}[e^{iz \cdot X}]$ (where $i = \sqrt{-1}$) and can be expanded as a power series whose terms are the moments of D

$$\widehat{D}(X) = \sum_{j=0}^{\infty} \frac{\mathbb{E}_{z \sim D}[(z \cdot X)^n]}{n!} i^n \,.$$

We can also invert the characteristic function to translate from the moments back to the actual density function. For our purposes we will want to work with the Hermite moments instead, because they can be robustly estimated using existing techniques (Kane, 2021; Liu and Moitra, 2021). It turns out that there is an analog of the relationship between moments and characteristic functions, but for Hermite moments instead. Specifically we define the adjusted characteristic function (for details see Definition 42) and show that the terms in the power series expansion of the adjusted characteristic function are exactly the Hermite moments (see Corollary 44). The key is we can give quantitative estimates for inverting the adjusted characteristic function that allow us to relate the size of terms in the power series (which are Hermite moments) to the L^1 norm of the density function.

This gives us some relation between the Hermite moments and the TV distance but it is still far from strong observability. In particular, the power series has infinitely many terms, but in order to prove strong observability, we must restrict to a constant ($O_k(1)$) number of test functions. The key is to prove that, for mixtures of Gaussians, the Hermite moments satisfy a recurrence relation of order $O_k(1)$ and further that this recurrence has bounded coefficients. For context, the moments of a single Gaussian satisfy a simple recurrence so it is reasonable to expect that the moments of a mixture satisfy a higher-order recurrence. This framework of working with Hermite moments through their recurrence relations (instead of through their parameters as in previous works) is crucial to circumventing the $\epsilon^{O(1/k)}$ barrier. This is only a sketch and the full proof has additional complexities.

However strong observability alone does not get us anywhere because, after robustly estimating the Hermite moments, we would still need to solve a large system of polynomial equations to find a good (proper) estimate. In fact this system does not appear to have any useful structure that can be exploited algorithmically, in part because it has many disconnected solutions due to the failure of robust parameter learning. Instead we circumvent this obstacle by showing how to solve a relaxed version of the polynomial system that corresponds to allowing ourselves to output a Gaussian mixture model whose mixing weights are low degree polynomials – i.e. it has the form

$$Q_1(x)G_1 + \cdots + Q_k(x)G_k$$

where $Q_1, \ldots, Q_k$ are low-degree polynomials that are nonnegative everywhere. We call this a mixture of polynomial Gaussians (MPG). As it turns out, our strong observability result (Theorem 50), that being close in Hermite moments implies closeness in total variation distance, extends to MPGs as well. We emphasize that this is only a high-level description of the proof, and there are many subtleties such as the crucial fact that the degrees of the polynomials $Q_1, \ldots, Q_k$ does not need to grow with m, the number of Hermite moments that we want to match, which is essential in order to make our strategy work. We give a detailed technical overview in Section 2.

1.3. Other Related Work

There has also been a line of work (Ashtiani et al., 2018, 2020) towards achieving optimal sample complexity for robustly learning mixtures of Gaussians. However, these works are not algorithmic i.e. their algorithms are based on discretization and brute-force search and run in exponential time. In our work, sample complexity is not our primary concern (as long as it is polynomial) and in order to develop a computationally efficient learning algorithm we will need entirely new techniques.

2. Technical Overview

We now present our technical overview. Due to space constraints, the formal theorems and proofs are deferred to the appendix (Section C and after).

2.1. Problem Setup

We use $N(\mu, \Sigma)$ to denote a Gaussian with mean μ and covariance Σ. We use $d_{\mathsf{TV}}(\mathcal{D}, \mathcal{D}')$ to denote the total variation distance between two distributions $\mathcal{D}, \mathcal{D}'$. When there is no ambiguity, we will slightly abuse notation and for a distribution $\mathcal{D}$ on $\mathbb{R}^d$, we use $\mathcal{D}(x)$ to denote the density function of $\mathcal{D}$ at x.

We use the following shorthand notation. X denotes a d-tuple of variables $(X_1, \dots, X_d)$. For a vector $\mu \in \mathbb{R}^d$ and matrix $\Sigma \in \mathbb{R}^{d \times d}$ we set $\mu(X) = \mu^T X$ and $\Sigma(X) = X^T \Sigma X$.

We begin by formally defining the problem that we will study. First we define the contamination model. This is a standard definition from robust learning (see e.g. Diakonikolas et al. (2020a)).

Definition 7 (Strong Contamination Model) *We say that a set of vectors $Y_1, \dots, Y_n$ is an ϵ-corrupted sample from a distribution $\mathcal{D}$ over $\mathbb{R}^d$ if it is generated as follows. First $X_1, \dots, X_n$ are sampled i.i.d. from $\mathcal{D}$. Then a (malicious, computationally unbounded) adversary observes $X_1, \dots, X_n$ and replaces up to ϵn of them with any vectors it chooses. The adversary may then reorder the vectors arbitrarily and output them as $Y_1, \dots, Y_n$*

We study the following problem. There is an unknown mixture of Gaussians $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ where $G_i = N(\mu_i, \Sigma_i)$. We receive an ϵ-corrupted sample $Y_1, \dots, Y_n$ from $\mathcal{M}$ where $n = \mathrm{poly}(d/\epsilon)$ (we treat k as a constant). The goal is to output a density function of a distribution, say f, such that $d_{\mathsf{TV}}(f, \mathcal{M}) \leq \widetilde{O}(\epsilon)$.

In our main result, Theorem 72, we give an algorithm that computes such a function f of the form

$$f(x) = Q_1(x)\overline{G}_1(x) + \dots + Q_k(x)\overline{G}_k(x)$$

where $Q_1, \dots, Q_k$ are polynomials of constant (possibly depending on k) degree that are nonnegative everywhere and $\overline{G}_1, \dots, \overline{G}_k$ are Gaussians. We call such functions mixtures of polynomial Gaussians (MPG) distributions for short (see Definition 23).

Throughout our paper, we will assume that all of the Gaussians that we consider have variance at least $\mathrm{poly}(\epsilon/d)$ and at most $\mathrm{poly}(d/\epsilon)$ in all directions i.e. they are not too flat. This implies that their covariance matrices are invertible so we may write expressions such as Σ_i^{-1}.

Remark 8 *Our main results for nearly optimal density estimation require a stronger assumption that the variances are between $\mathrm{poly}(\log 1/\epsilon)^{-1}$ and $\mathrm{poly}(\log 1/\epsilon)$ in each direction. However, en route to these results, we first prove a few simple generalizations of the the results in Liu and Moitra (2021) and these results hold under the same assumptions as in Liu and Moitra (2021) i.e. components have variance between $\mathrm{poly}(\epsilon/d)$ and $\mathrm{poly}(d/\epsilon)$ in all directions.*

We will also assume that the w_i are at least A^{-1} for some constant A. While a lower bound on the mixing weights is not technically necessary for density estimation, we need such an assumption in our paper because we need to first run a parameter estimation algorithm (see Liu and Moitra (2021)) to obtain rough estimates for all of the components.

Throughout this paper, we treat k, A as constants – i.e. A could be any function of k – and when we say polynomial, the exponent may depend on these parameters. We are primarily interested in dependence on ϵ and d (the dimension of the space).

2.2. Main Ideas

2.2.1. Regular Form Mixtures

We will first consider the case when the components of the mixture are in a convenient form, which we call regular form, meaning that all of the components are not too far from each other and not too poorly conditioned.

Definition 9 (Informal, see Definition 24)) *We say a mixture of Gaussians $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ is in regular-form if all of the components can be written in the form $G_j = N(\mu_j, I + \Sigma_j)$ where $\|\mu_j\|, \|\Sigma_j\|_2 \leq \text{poly}(\log 1/\epsilon)$ and $\text{poly}(\log 1/\epsilon)^{-1} I \leq I + \Sigma_j \leq \text{poly}(\log 1/\epsilon) I$*

In the appendix (Section B), we sketch how we reduce from a general mixture to a mixture in regular form. As mentioned previously, the key ingredient in our algorithm is an observability statement, that we have a family of test functions such that if two mixtures are close on this family of test functions, then they must be close in TV distance. Our learning algorithm will then work by measuring these test functions using the samples and then solving for a distribution that matches these test function measurements.

For many learning problems, such as learning mixtures of Gaussians in the non-robust setting (Moitra and Valiant, 2010; Kalai et al., 2010), using low-degree moments as the set of test functions suffices. However, for robustly learning mixtures of Gaussians, using the standard moments would lose factors of $\text{poly}(d)$ in the error guarantee. Similar to previous papers on robustly learning mixtures of Gaussians (Kane, 2021; Liu and Moitra, 2021), we use the family of low-degree Hermite moments as our test functions. Of course, as mentioned previously, there are still many additional obstacles to proving strong observability and circumventing the $\epsilon^{1/k}$ barrier to parameter learning. First, we make a few definitions. See Section D.1 for more details.

Definition 10 *Let $\mathcal{H}_m(x)$ denote the univariate Hermite polynomials, e.g. $\mathcal{H}_2(x) = x^2 - 1, \mathcal{H}_3(x) = x^3 - 3x$. Let $\mathcal{H}_m(x, y^2)$ be the homogenized Hermite polynomials e.g. $\mathcal{H}_2(x, y^2) = x^2 - y^2, \mathcal{H}_3(x, y^2) = x^3 - 3xy^2$.*

Definition 11 (Multivariate Hermite Polynomials) *Let $H_m(X, z)$ be a formal polynomial in variables $X = X_1, \dots, X_d$ whose coefficients are polynomials in d variables $z_1, \dots, z_d$ that is given by $H_m(X, z) = \mathcal{H}_m(z_1 X_1 + \cdots + z_d X_d, X_1^2 + \cdots + X_d^2)$.*

Definition 12 (Hermite Moment Polynomials) *For a distribution D on $\mathbb{R}^d$, we let*

$$h_{m,D}(X) = \mathbb{E}_{(z_1,\dots,z_d)\sim D}[H_m(X, z)] .$$

We omit the subscript D when it is clear from context. We refer to $h_{m,D}(X)$ as the Hermite moment polynomials of D.

Remark 13 *We will use the term Hermite moment polynomial (instead of just Hermite moment) to emphasize the fact that we view $h_{m,D}(X)$ as a polynomial in the formal variables X. This polynomial representation will be useful for the machinery that we introduce later on for manipulating these Hermite moment polynomials.*

Remark 14 *If instead of $\mathbb{E}_{(z_1,\dots,z_d)\sim D}[H_m(X,z)]$ we had $\mathbb{E}_{(z_1,\dots,z_d)\sim D}[(z_1X_1+\cdots+z_dX_d)^m]$ then we would get the standard moments.*

The Hermite moment polynomials turn out to be a particularly nice object to work with because we can work with their L^2 norm (after reorganizing the coefficients into a vector) without losing dimension dependent factors [2]. The key observability result in our proof (Theorem 50) implies that for regular-form mixtures, the distance between Hermite moment polynomials (in terms of L^2 norm of coefficients) is equivalent to TV distance up to a $\mathrm{poly}(\log 1/\epsilon)$ factor.

Algorithm Summary: We now summarize our algorithm for learning regular-form mixtures. The main parts are

1. **Strong Observability through Hermite Moment Polynomials:** we prove that for two mixtures of k Gaussians, and more generally mixtures of k polynomial Gaussians, if their first $O_k(1)$ Hermite moment polynomials are ϵ-close in coefficient L^2 distance, then the two mixtures are $\widetilde{O}(\epsilon)$-close in TV distance (Theorem 50).

2. **Estimate the Hermite Moment Polynomials Optimally:** we estimate the Hermite moment polynomials of the mixture to optimal accuracy (Theorem 65)

3. **Compute Rough Component Estimates:** we compute $\epsilon^{\Omega_k(1)}$-accurate estimates for all of the components (Theorem 73)

4. **Estimate Density Function Optimally:** we bootstrap the rough component estimates using the Hermite moment polynomial estimates to compute the density function of the mixture to optimal accuracy (Theorem 69)

The next figure shows how the parts fit together in our algorithm. We focus on parts $1, 2, 4$ because part 3 follows easily from the results in Liu and Moitra (2021).

Strong Observability through Hermite Moment Polynomials: To help build intuition, here we will sketch a proof of observability via Hermite moment polynomials in the infinite limit, i.e. if two mixtures match exactly on their first $O_k(1)$ Hermite moment polynomials, then they must be exactly the same. For simplicity, in this discussion, we will restrict ourselves to mixtures of Gaussians. In our analysis later on, we will need observability for mixtures of polynomial Gaussians because the density function that we output is in this more general class.

Here, we sketch a proof of the following (informal) theorem. The full version is in Theorem 50.

Theorem 15 (Informal) *If two regular-form mixtures of Gaussians $\mathcal{M} = w_1G_1+\cdots+w_kG_k$ and $\mathcal{M}' = w_1'G_1'+\cdots+w_k'G_k'$ match on their first $O_k(1)$ Hermite moment polynomials then $\mathcal{M} = \mathcal{M}'$.*

[2] For, say, standard moments, we would instead need to work with the tensor injective norm.

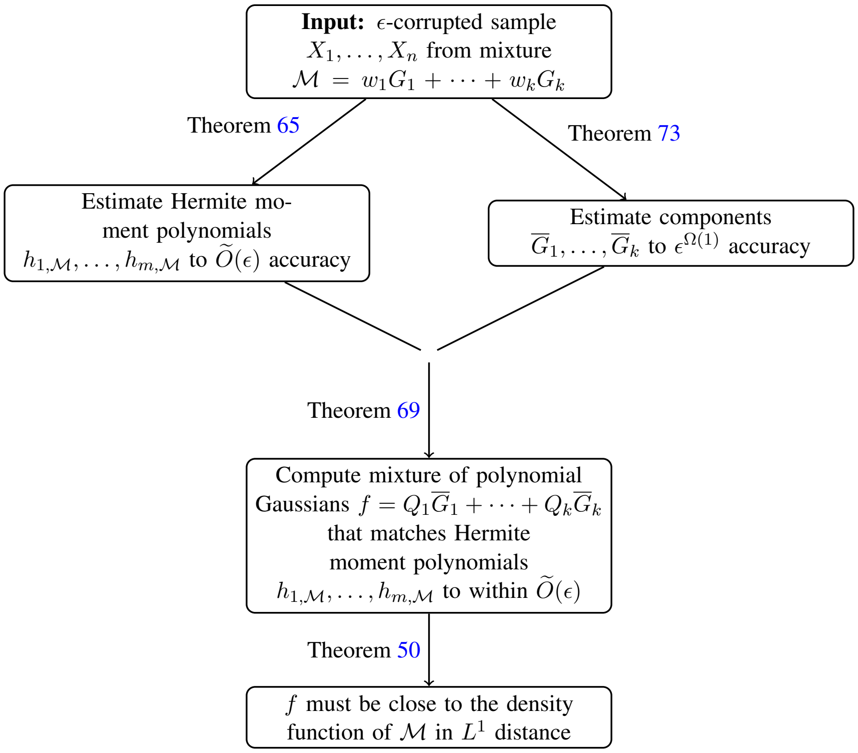
Input: ϵ-corrupted sample $X_1, \dots, X_n$ from mixture $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$
Theorem 65
Theorem 73
Estimate Hermite moment polynomials $h_{1,\mathcal{M}}, \dots, h_{m,\mathcal{M}}$ to $\widetilde{O}(\epsilon)$ accuracy
Estimate components $\overline{G}_1, \dots, \overline{G}_k$ to $\epsilon^{\Omega(1)}$ accuracy
Theorem 69
Compute mixture of polynomial Gaussians $f = Q_1\overline{G}_1 + \dots + Q_k\overline{G}_k$ that matches Hermite moment polynomials $h_{1,\mathcal{M}}, \dots, h_{m,\mathcal{M}}$ to within $\widetilde{O}(\epsilon)$
Theorem 50
f must be close to the density function of $\mathcal{M}$ in L^1 distance

Figure 1: Overview of our algorithm for regular-form mixtures

A key ingredient will be understanding recurrence relations between Hermite moment polynomials. To obtain these recurrence relations, we write down a generating function for the Hermite moment polynomials and then manipulate this generating function using differential operators. By writing down a differential operator that annihilates the generating function, we then obtain the coefficients of a recurrence relation that the Hermite moment polynomials must satisfy. First, we have the following identity.

Claim 1 (See Corollary 40) *Let $\mathcal{M} = w_1 G_1 + \dots w_k G_k$ be a mixture of Gaussians where $G_j = N(\mu_j, I + \Sigma_j)$. Then*

$$\sum_{m=0}^{\infty} \frac{1}{m!} \cdot h_m(X) y^m = w_1 e^{\mu_1(X) + \frac{1}{2}\Sigma_1(X) y^2} + \dots + w_k e^{\mu_k(X) + \frac{1}{2}\Sigma_k(X) y^2} \tag{1}$$

where $h_m(X)$ are the Hermite moment polynomials of the mixture $\mathcal{M}$.

Let $f(y)$ be the function on the right hand side of (1), viewed as a function of y with formal variables X. For each $j \in [k]$, consider the differential operator $\mathcal{D}_j = \partial - (\mu_j(X) + \Sigma_j(X) y)$ where the partial derivative is taken with respect to y. Then if we let $\mathcal{D} = \mathcal{D}_k^{2^{k-1}} \mathcal{D}_{k-1}^{2^{k-2}} \cdots \mathcal{D}_1$, we can verify that $\mathcal{D}(f(y)) = 0$ (see Section D.5).

On the other hand, by using the product rule, we can expand the differential operator $\mathcal{D}$ in the form

$$\mathcal{D} = Q_{2^k - 1}(X, y) \partial^{2^k - 1} + Q_{2^k - 2}(X, y) \partial^{2^k - 2} + \dots + Q_0(X, y)$$

where $Q_0, \dots, Q_{2^k-1}$ are polynomials in y whose coefficients are polynomials in the formal variables X. It is immediate to verify that Q_j has degree at most 2^k-1-j in y so for all $0 \leq j \leq 2^k-1$ we can write

$$Q_j(X,y) = R_{j,2^k-1-j}(X)y^{2^k-1-j} + \dots + R_{j,0}(X)$$

for some polynomials $R_{j,0}(X), \dots, R_{j,2^k-1-j}(X)$.

Now consider what happens when we apply $\mathcal{D}$ to the left hand side of (1). We will get a power series in y whose coefficients are polynomials in X. The coefficient of y^a will be

$$\sum_{j=0}^{2^k-1} \sum_{l=0}^{2^k-1-j} \frac{h_{a+j-l}(X)R_{j,l}(X)}{(a-l)!}$$

and since $\mathcal{D}(f(y)) = 0$ we get the following conclusion.

Claim 2 *Let $\mathcal{M} = w_1G_1 + \dots w_kG_k$ be a mixture of Gaussians where $G_j = N(\mu_j, I + \Sigma_j)$. Then there are polynomials $R_{j,l}(X)$ such that for all a, the Hermite moment polynomials of $\mathcal{M}$ satisfy*

$$\sum_{j=0}^{2^k-1} \sum_{l=0}^{2^k-1-j} \frac{h_{a+j-l}(X)R_{j,l}(X)}{(a-l)!} = 0\,. \tag{2}$$

This means that the Hermite moment polynomials satisfy a recurrence of order $O_k(1)$. It is straightforward to extend the above argument to the difference of two mixtures say $\mathcal{M} = w_1G_1 + \dots + w_kG_k$ and $\mathcal{M}' = w_1'G_1' + \dots + w_k'G_k'$ and we deduce that the polynomials $h_{m,\mathcal{M}}(X) - h_{m,\mathcal{M}'}(X)$ must satisfy a similar recurrence of order $O_k(1)$. Thus, if the first $O_k(1)$ Hermite moment polynomials of two mixtures are the same, then all of their Hermite moment polynomials must be the same. Note that this statement suffices for the proof of observability in the infinite limit, but in the full analysis, we need several additional steps to prove quantitative bounds relating the distance between higher-degree Hermite moment polynomials to the distance between the first $O_k(1)$ Hermite moment polynomials.

The argument above implies that

$$\sum_{m=0}^{\infty} \frac{1}{m!} \cdot h_{m,\mathcal{M}}(X)y^m = \sum_{m=0}^{\infty} \frac{1}{m!} \cdot h_{m,\mathcal{M}'}(X)y^m\,. \tag{3}$$

It remains to show how to transform from these generating functions back to the original distributions. This can be done through the adjusted characteristic function. We define

Definition 16 (Adjusted Characteristic Function) *For a distribution D on $\mathbb{R}^d$, we define its adjusted characteristic function $\widetilde{D} : \mathbb{R}^d \to \mathbb{C}$ as $\widetilde{D}(X) = \mathbb{E}_{z\sim D}\left[e^{iz\cdot X + \frac{1}{2}\|X\|^2}\right]$ where $i = \sqrt{-1}$.*

It suffices to note that

Claim 3 (Same as Claim 14) *Let D be a distribution on $\mathbb{R}^d$. Then $\widetilde{D}(X) = \sum_{m=0}^{\infty} \frac{i^m}{m!} h_{m,D}(X)$.*

Thus, plugging $y = i$ into (3), we get that $\widetilde{\mathcal{M}}(X) = \widetilde{\mathcal{M}'}(X)$. However, note that the adjusted characteristic function is an invertible transformation (since we can multiply by $e^{-\frac{1}{2}\|X\|^2}$ and then invert the characteristic function) so actually $\mathcal{M}(X) = \mathcal{M}'(X)$, completing the proof of observability in the infinite limit.

There are several additional technical ingredients that are necessary to go from observability in the infinite limit to quantitatively strong observability. The main one is that we need to bound the coefficients of the polynomials $R_{j,l}(X)$. It is not difficult to obtain sufficiently tight bounds when $\|\mu_j\|, \|\Sigma_j\|_2$ are all sufficiently small (smaller than some constant depending on k). To reduce to this case, we need to do some additional work (see Section E.3). Also, we will need quantitative bounds on inverting the adjusted characteristic function. Such bounds are obtained in Section D.2.

Estimate the Hermite Moment Polynomials Optimally: Here, our goal is to obtain estimates $h'_m(X)$ for the first $O_k(1)$ Hermite moment polynomials. We sketch a proof of the following (informal) theorem. The full version is in Theorem 65.

Theorem 17 (Informal) *Given an ϵ-corrupted sample from a regular-form mixture of Gaussians $\mathcal{M} = w_1G_1 + \cdots + w_kG_k$, we can compute estimates $h'_m(X)$ such that $\|v(h_m(X) - h'_m(X))\| \leq \widetilde{O}(\epsilon)$ where $h_m(X)$ are the true Hermite moment polynomials of the mixture and $v(\cdot)$ denotes converting a polynomial to a vector of coefficients (we then measure the L^2 norm of this vector).*

Note that $h_m(X)$ is the mean of the distribution of $H_m(X, z)$ for $z \sim \mathcal{M}$ so estimating $h_m(X)$ is a robust mean estimation task. Previous papers (Liu and Moitra, 2021; Kane, 2021) estimate $h_m(X)$ up to accuracy $\widetilde{O}(\sqrt{\epsilon})$ by upper bounding the spectral norm of the covariance matrix and using standard results from robust mean estimation. However achieving $\widetilde{O}(\epsilon)$ is significantly more difficult.

It can be shown, via hypercontractivity, that the distribution of $H_m(X, z)$ exhibits exponential tail decay (see Lemma 61). However, this alone is not enough to robustly estimate the mean to within $\widetilde{O}(\epsilon)$ in a computationally efficient manner. Existing results achieving optimal accuracy e.g. Diakonikolas et al. (2020b) require known covariance or some additional moment structure (such as in the case of a single Gaussian). Furthermore, there is evidence suggesting that achieving optimal accuracy for general sub-Gaussian distributions may be computationally hard (Hopkins and Li, 2019).

To circumvent these barriers, we leverage the structure of the moments of the distribution of $H_m(X, z)$. Roughly speaking, we write the covariances of the distributions of $H_0(X, z), \dots, H_m(X, z)$ in terms of the Hermite moment polynomials $h_0(X), \dots, h_m(X)$ (which are the means of the respective distributions). Thus, we can estimate the means, compute estimates for the covariances and then use our covariance estimates to refine our estimates of the means and keep iterating. This is similar to how algorithms for robustly learning a single Gaussian use the relation between its covariance and its fourth moment tensor. Of course, the moment structure of the distribution of $H_m(X, z)$ is significantly more complex so the analysis will be more involved.

Note that if the covariance of $H_m(X, z)$ were known, then we would be able to estimate the mean of the distribution to $\widetilde{O}(\epsilon)$ accuracy using standard techniques (e.g. Diakonikolas et al. (2020b)). The first important observation is that the covariance of the distribution of $H_m(X, z)$ can be written in terms of the first $2m$ Hermite moment polynomials $h_0(X), \dots, h_{2m}(X)$. Next, if m is sufficiently large in terms of k, then the polynomials $h_{m+1}(X), \dots, h_{2m}(X)$ can be computed in terms of $h_0(X), \dots, h_m(X)$ via the recurrence in (2). Since we do not know the actual recurrence, we can solve for the coefficients in the recurrence using $h_0(X), \dots, h_m(X)$ and then use these coefficients to extend the recurrence and compute $h_{m+1}(X), \dots, h_{2m}(X)$. With this insight, we can build an iterative algorithm that repeatedly refines our estimates. Using an upper bound on the covariance (same as in Liu and Moitra (2021); Kane (2021)), we can ensure that our initial estimates are $\widetilde{O}(\sqrt{\epsilon})$-accurate. Then, by repeatedly running the above, we can refine these estimates to $\widetilde{O}(\epsilon)$-accuracy. (see Section G.3).

Compute Rough Component Estimates: Here, our goal is to prove the following (informal) theorem. See Theorem 73 for the full version.

Theorem 18 (Informal) *Given an ϵ-corrupted sample from a mixture $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$, we can compute estimates $\overline{G}_1, \dots, \overline{G}_k$ for the components such that $d_{\mathsf{TV}}(G_j, \overline{G}_j) \leq \epsilon^{\Omega_k(1)}$.*

This theorem follows from a simple modification to the techniques in Liu and Moitra (2021). Note that the only difference is that the main theorem in Liu and Moitra (2021) assumes that the components of the mixture are not too close in TV distance. However, this assumption can be removed by essentially merging components and treating them as one if they are $\epsilon^{\Omega(1)}$-close.

Estimate Density Function Optimally: So far, we showed that the first $O_k(1)$ Hermite moment polynomials suffice to determine a mixture of Gaussians (and more generally a mixture of polynomial Gaussians). We then showed how to estimate these Hermite moment polynomials to optimal accuracy. The last step is to compute a mixture of polynomial Gaussians that matches these Hermite moment polynomials. To do this, we will take the rough estimates of the components from the previous step and then multiply them by appropriate polynomials. We sketch a proof of the following (informal) theorem. The full version is in Theorem 69.

Theorem 19 (Informal) *Let $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ be a mixture of Gaussians in regular form. Assume we are given estimates $\overline{G}_1, \dots, \overline{G}_k$ for the components such that for all j, $d_{\mathsf{TV}}(G_j, \overline{G}_j) \leq \epsilon^{\Omega_k(1)}$. Then we can compute a distribution $f(x) = Q_1(x)\overline{G}_1(x) + \dots + Q_k(x)\overline{G}_k(x)$ where $Q_1, \dots, Q_k$ are polynomials of degree $C = O_k(1)$ such that the first $m = O_{k,C}(1)$ Hermite moment polynomials of f match those of $\mathcal{M}$ up to $\widetilde{O}(\epsilon)$ accuracy.*

It is crucial to note that in order to match m Hermite moment polynomials, the degree of the polynomials $Q_1, \dots, Q_k$ that we need *does not grow with m*. In other words, we first fix the degree C of the polynomials $Q_1, \dots, Q_k$. We then argue that for any m, we can match the first m Hermite moment polynomials using polynomials $Q_1, \dots, Q_k$ of degree C. The only place that m shows up is in the accuracy i.e. the $\widetilde{O}(\epsilon)$ hides a factor of the form $\epsilon(\log 1/\epsilon)^m$.

The reason that we need to fix C first and then choose m sufficiently large in terms of C is that the observability result, Theorem 50, only works when two distributions match on their first m Hermite moment polynomials for m much larger than C, k.

We now sketch how we actually compute the polynomials $Q_1, \dots, Q_k$. For simplicity, we will first consider a single Gaussian $G = N(\mu, I + \Sigma)$ as this already will reveal the key intuitions. Assume that we are given an estimate of G, say $\overline{G} = N(\widetilde{\mu}, I + \widetilde{\Sigma})$ with $d_{\mathsf{TV}}(G, \overline{G}) \leq \epsilon^c$ for some constant c. Recall Claim 1 which implies $\sum_{m=0}^{\infty} \frac{1}{m!} \cdot h_{m,G}(X) y^m = e^{\mu(X)y + \frac{1}{2}\Sigma(X)y^2}$. Now consider the generating function

$$e^{\mu(X)y + \frac{1}{2}\Sigma(X)y^2} = e^{(\mu(X) - \widetilde{\mu}(X))y + \frac{1}{2}(\Sigma(X) - \widetilde{\Sigma}(X))y^2} e^{\widetilde{\mu}(X)y + \frac{1}{2}\widetilde{\Sigma}(X)y^2} .$$

Since $d_{\mathsf{TV}}(G, \overline{G}) \leq \epsilon^c$, we can show that $\|\mu - \widetilde{\mu}\|, \left\|\Sigma - \widetilde{\Sigma}\right\|_2 \leq \epsilon^{\Omega(c)}$. Now consider the power series expansion of

$$e^{(\mu(X) - \widetilde{\mu}(X))y + \frac{1}{2}(\Sigma(X) - \widetilde{\Sigma}(X))y^2} = \sum_{m=0}^{\infty} \frac{\left((\mu(X) - \widetilde{\mu}(X))y + \frac{1}{2}(\Sigma(X) - \widetilde{\Sigma}(X))y^2\right)^m}{m!} \,,$$

We can expand each term on the RHS using the binomial theorem. The key observation is that since $\|\mu - \widetilde{\mu}\|, \left\|\Sigma - \widetilde{\Sigma}\right\|_2 \leq \epsilon^{\Omega(c)}$, whenever we multiply more than $O(1/c)$ terms of the form $(\mu(X) - \widetilde{\mu}(X))$ or $(\Sigma(X) - \widetilde{\Sigma}(X))$ together, the result will have coefficient norm smaller than ϵ. Thus, we can essentially drop all but the first $O(1/c)$ terms in the power series expansion i.e.

$$e^{(\mu(X) - \widetilde{\mu}(X))y + \frac{1}{2}(\Sigma(X) - \widetilde{\Sigma}(X))y^2} \sim P(X, y)$$

where P has degree at most $O(1/c)$ in y and X. Thus, $e^{\mu(X)y + \frac{1}{2}\Sigma(X)y^2} \sim P(X, y)e^{\widetilde{\mu}(X)y + \frac{1}{2}\widetilde{\Sigma}(X)y^2}$. It remains to plug in $y = i$, multiply by $e^{-\frac{1}{2}\|X\|^2}$ and invert the characteristic function (recall Claim 3). We can then verify that the resulting function will be of the form $Q(x)\overline{G}(x)$ where Q has degree at most $O(1/c)$.

The above intuition roughly says that G can be approximated by $\overline{G}$ times a polynomial of degree $O(1/c)$ as long as $d_{\mathsf{TV}}(G, \overline{G}) \leq \epsilon^c$. Thus, the mixture of Gaussians $\mathcal{M} - w_1 G_1 + \cdots + w_k G_k$ can be approximated by a function of the form $f(x) = Q_1(x)\overline{G}_1(x) + \cdots + Q_k(x)\overline{G}_k(x)$ for some constant-degree polynomials $Q_1, \ldots, Q_k$. It remains to show how to compute the polynomials $Q_1, \ldots, Q_k$. To solve for $Q_1, \ldots, Q_k$, it suffices to note that the Hermite moment polynomials of $f(x)$ are linear forms in the coefficients of $Q_1, \ldots, Q_k$. Thus, since we have estimates for the Hermite moment polynomials of the true mixture $\mathcal{M}$, it suffices to solve a linear system for the coefficients of $Q_1, \ldots, Q_k$.

Acknowledgments

Allen Liu was supported in part by an NSF Graduate Research Fellowship, a Fannie and John Hertz Foundation Fellowship and Ankur Moitra's NSF CAREER Award CCF-1453261 and NSF Large CCF1565235. Ankur Moitra was supported in part by a Microsoft Trustworthy AI Grant, NSF CAREER Award CCF-1453261, NSF Large CCF1565235, a David and Lucile Packard Fellowship and an ONR Young Investigator Award.

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Lavrentin M Arutyunyan, Egor D Kosov, et al. Deviation of polynomials from their expectations and isoperimetry. *Bernoulli*, 24(3):2043–2063, 2018.

Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. Sample-efficient learning of mixtures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67:1–42, 2020.

Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *arXiv preprint arXiv:2005.02970*, 2020.

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. Robustly learning mixtures of k arbitrary gaussians, 2020.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Online and distribution-free robustness: Regression and contextual bandits with huber contamination. *arXiv preprint arXiv:2010.04157*, 2020.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019. URL http://arxiv.org/abs/1911.05911.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org, 2017a.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017b.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019a.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606, 2019b.

Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020a.

Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/13ec9935e17e00bed6ec8f06230e33a9-Abstract.html.

Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pages 761–770, 2015.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760, 2015.

Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. ACM, 2018.

Samuel B. Hopkins and Jerry Li. How hard is robust mean estimation? In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 1649–1682. PMLR, 2019. URL http://proceedings.mlr.press/v99/hopkins19a.html.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

Daniel M. Kane. Robust learning of mixtures of gaussians. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 1246–1258. SIAM, 2021. doi: 10.1137/1.9781611976465.76. URL https://doi.org/10.1137/1.9781611976465.76.

Manuel Kauers, Ryan O'Donnell, Li-Yang Tan, and Yuan Zhou. Hypercontractive inequalities via sos, and the frankl–rödl graph. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1644–1658. SIAM, 2014.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430, 2018.

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

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

Jerry Zheng Li. *Principled approaches to robust machine learning and beyond*. PhD thesis, Massachusetts Institute of Technology, 2018.

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 *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Jacob Steinhardt. *Robust Learning: Information Theory and Algorithms*. PhD thesis, Stanford University, 2018.

Henry Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, 1961.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

Appendix A. Comparison of Observability and Identifiability

It will be helpful to compare observability with the more familiar concept of identifiability, which is usually thought of as the crucial ingredient in parameter learning algorithms. Recall the definition of identifiability.

Definition 20 (Identifiability, informal) *Given a family of distributions $\mathcal{F}$ parameterized by parameters θ, we say that the family $\mathcal{F}$ is identifiable if any two distributions $\mathcal{F}(\theta), \mathcal{F}(\theta') \in \mathcal{F}$ that are close in TV they must also be close in terms of their parameters (for some appropriate parameter distance).*

In the case of parameter distance for GMMs, usually we require that there be a matching between the components in $\mathcal{M}$ and those in $\mathcal{M}'$ so that across the matching the components are close in TV and have similar mixing weights.

However identifiability is just not the right notion to use for density estimation, at least when it comes to achieving nearly optimal robustness guarantees. The issue is that the relationship between component-wise distance and TV distance is quantitatively too weak. There are explicit constructions of GMMs $\mathcal{M}$ and $\mathcal{M}'$ that are ϵ-close in TV distance but where all the components in both mixtures are all separated by at least $\epsilon^{O(1/k)}$ (see Moitra and Valiant (2010); Hardt and Price (2015)) in TV. The key point is that the mixtures are so close to each other that we cannot distinguish between them in the setting where an ϵ-fraction of our samples can be arbitrarily corrupted. Hence we have:

Proposition 21 *When an ϵ-fraction of the samples are arbitrarily corrupted, it is not information-theoretically possible to learn the components of a GMM to accuracy better than $\epsilon^{O(1/k)}$.*

This is a serious issue because it means that many of the standard techniques for learning mixture models that work by learning individual components are trying to do too much and will get stuck at the above barrier. In particular, while we are using the same family of test functions as Liu and Moitra (2021), the techniques in Liu and Moitra (2021) rely on trying to isolate the parameters of each of the components and thus will run into the $\epsilon^{O(1/k)}$ barrier.

Appendix B. Overview of Reducing to Regular Form

For general mixtures, our algorithm has one additional step where we need to cluster the mixture into submixtures and then place each submixture in regular-form.

To do this, we can use Theorem 73 to obtain rough estimates for all of the components. We then cluster the samples into subsamples by assigning each sample to the estimated component that assigns it the highest likelihood. While this clustering will not classify all of the samples "correctly" (e.g. if components overlap), we combine the subsamples into submixtures and argue that for some recombination the following two conditions hold:

- The clustering into submixtures is accurate to within $\widetilde{O}(\epsilon)$ accuracy (see Lemma 82)

- For each submixture, we can apply a linear transformation to place it in regular form

Thus, once we find this recombination (say by enumerating over all possible recombinations), we have reduced the problem to learning mixtures of Gaussians in regular form.

Appendix C. Notation and Preliminaries

C.1. Basic Definitions

We now introduce some terminology that we will use throughout the paper.

Definition 22 *We say a function $f(x) : \mathbb{R}^d \to \mathbb{R}$ is a degree-m polynomial Gaussian if it can be written in the form*

$$f(x) = Q(x)G(x)$$

where $G(x)$ is the probability density function of a Gaussian and Q is a polynomial in d variables of degree at most m.

Definition 23 *We say a function $f(x) : \mathbb{R}^d \to \mathbb{R}$ is a degree-m mixture of polynomial Gaussians (MPG) if*

$$f(x) = Q_1(x)G_1(x) + \cdots + Q_k(x)G_k(x)$$

where $G_1(x), \dots, G_k(x)$ are the probability density functions of Gaussians and $Q_1, \dots, Q_k$ are polynomials in d variables of degree at most m. If the polynomials $Q_1, \dots, Q_k$ are all nonnegative for any $x \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} f(x)dx = 1$, then we say f is a degree-m MPG distribution.

We will often need to work with mixtures of Gaussians whose components are in a specific form, which we call regular-form.

Definition 24 *We say a set of Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$ is in (α, β)-regular form if the following holds:*

- *For all j, $\|\mu_j\| \leq \alpha$*
- *For all j, $\|\Sigma_j\|_2 \leq \alpha$*
- *For all j, $\beta^{-1} I \leq I + \Sigma_j \leq \beta I$*

We will sometimes need an additional conditions that there exists some j such that

$$\|\mu_j\| + \|\Sigma_j\|_2 \leq \gamma \,.$$

If this additional condition holds, we say that the set of Gaussians is in (α, β, γ)-regular form.

Definition 25 *We say that a mixture of Gaussians $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ is in (α, β) (respectively (α, β, γ)) regular-form if the set of components $\{G_1, \dots, G_k\}$ is in (α, β) (respectively (α, β, γ)) regular-form.*

Remark 26 *Generally, we will be interested in the regime where $\alpha, \beta \leq \text{poly}(\log 1/\epsilon)$ and γ is a sufficiently small constant (in terms of k).*

Definition 27 *Given a family of polynomials $\mathcal{S} = \{P_1, P_2, \dots\}$ in variables $X = (X_1, \dots, X_d)$, we say a polynomial $Q(X)$ is (A, B)-simple with respect to $\mathcal{S}$ for some parameters A, B if Q can be written as a linear combination of A terms where*

- *All coefficients in the linear combination have magnitude at most A*
- *Each term is a product of at most B polynomials from $\mathcal{S}$*

We will need the following standard fact (see e.g. Arutyunyan et al. (2018); Kauers et al. (2014)) that allows us to bound the tail decay of the distribution of a polynomial $f(x)$ where x is drawn from a Gaussian.

Claim 4 (Hypercontractivity) *Let f be a polynomial of degree m. Let $G = N(\mu, \Sigma)$ be a Gaussian in $\mathbb{R}^d$. There is a universal constant c such that for any even integer q,*

$$(\mathbb{E}_{x \sim G} |f(x)|^q) \leq (cq)^{mq} \left(\mathbb{E}_{x \sim G} |f(x)|^2\right)^{q/2} \,.$$

C.2. Tensors and Polynomials

We will now introduce notation and tools to deal with tensors and formal polynomials. We will need to translate between polynomials and their corresponding representations as tensors repeatedly in this paper.

Definition 28 *Let X denote the set of formal variables $(X_1, \dots, X_d)$. Then for a positive integer k, $X^{\otimes k}$ denotes the $\underbrace{d \times \cdots \times d}_{k}$ tensor*

$$\underbrace{(X_1, \dots, X_d) \otimes \cdots \otimes (X_1, \dots, X_d)}_{k}$$

Now we will define a canonical transformation between polynomials and tensors.

Definition 29 *For a homogeneous polynomial $f(X)$ of degree k in the d variables $X_1, \dots, X_d$ with real coefficients, define $T(f)$ to be the unique symmetric tensor with dimensions $\underbrace{d \times \cdots \times d}_{k}$ such that*

$$\langle T(f), X^{\otimes k} \rangle = f(X) \,.$$

We call $T(f)$ the coefficient tensor of f.

Definition 30 *For a homogeneous polynomial $f(X)$ in the d variables $X_1, \dots, X_d$ with real coefficients define $v(f)$ to be the vector obtained by flattening $T(f)$. We call $v(f)$ the coefficient vector of f.*

Definition 31 *For a polynomial (not necessarily homogeneous) $f(X, y)$, viewed as a polynomial in y whose coefficients are homogeneous polynomials in X (of not necessarily the same degree) i.e.*

$$f(X, y) = f_0(X) + f_1(X)y + \cdots + f_k(X)y^k$$

we define $v_y(f)$ to be the vector obtained by concatenating $v(f_m(X))$ for all m.

We will frequently consider expressions of the form $\|v(f)\|$ i.e. the L^2 norm of the coefficient vector.

Definition 32 *For a polynomial $f(X)$, we call $\|v(f)\|$ the coefficient norm of f.*

The first claim below gives us an upper bound on the coefficient norm of the product of two polynomials f and g in terms of the coefficient norms of f and g.

Claim 5 *Let f and g be two homogeneous polynomials in the variables $X = (X_1, \dots, X_d)$ of degree m_1, m_2 respectively. Then*

$$\|T(fg)\|_2 \leq \|T(f)\|_2 \, \|T(g)\|_2 \,.$$

Equivalently,

$$\|v(fg)\| \leq \|v(f)\| \, \|v(g)\|$$

Proof Note that $T(fg)$ can be written as an average, over all partitions of $[m_1 + m_2]$ into two sets S_1, S_2 of size m_1, m_2, of $T(f)_{S_1} \otimes T(g)_{S_2}$ where $T(f)_{S_1} \otimes T(g)_{S_2}$ is a $d^{\otimes(m_1+m_2)}$ tensor obtained by taking $T(f)$ in the dimensions indexed by S_1 and $T(g)$ in the dimensions indexed by S_2 and tensoring them together. It is clear that

$$\|T(f) \otimes T(g)\|_2 = \|T(f)\|_2 \, \|T(g)\|_2$$

so using the triangle inequality, we get the desired conclusion. ■

We also have a lower bound on $\|v(fg)\|$ that follows immediately from the results in Liu and Moitra (2021).

Claim 6 (Claim 3.18 in Liu and Moitra (2021)) *Let f, g be two homogeneous polynomials in the variables $X = (X_1, \dots, X_d)$ of degree at most m. Then*

$$\|v(fg)\| \geq \Omega_m(1) \, \|v(f)\| \, \|v(g)\| \; .$$

The next claim gives us an understanding of how linear transformations of the underlying variables $X = (X_1, \dots, X_d)$ affect the coefficient norm of a polynomial.

Claim 7 *Let f be a homogeneous polynomial in the variables $X = (X_1, \dots, X_d)$ of degree equal to m. Let Σ be a $d \times d$ matrix. Then*

$$\|v(f(\Sigma X))\| \leq \left(\|\Sigma\|_{op}\right)^m \|v(f(X))\| \; .$$

Proof Note that

$$v(f(\Sigma X)) = \left(\underbrace{\Sigma \otimes \dots \otimes \Sigma}_{m}\right) v(f(X)) \, .$$

where $\otimes$ in the above denotes the Kronecker product. Also,

$$\left\| \underbrace{\Sigma \otimes \dots \otimes \Sigma}_{m} \right\|_{\mathrm{op}} = \left(\|\Sigma\|_{\mathrm{op}}\right)^m$$

and now we immediately get the desired inequality. ■

C.3. Tensors and Polynomials with Multiple Sets of Variables

We will need a few additional definitions dealing with polynomials and tensors involving multiple sets of variables, say $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)}), \dots, X^{(k)} = (X_1^{(k)}, \dots, X_d^{(k)})$. We first prove the following property.

Claim 8 *Let k be a positive integer and consider k distinct sets of d formal variables, say $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)}), \dots, X^{(k)} = (X_1^{(k)}, \dots, X_d^{(k)})$. Let A be the tensor of polynomials defined as follows:*

$$A = \mathit{flat}\left(\left(X^{(1)}\right)^{\otimes m_1}\right) \otimes \cdots \otimes \mathit{flat}\left(\left(X^{(k)}\right)^{\otimes m_k}\right) .$$

Note that A is an order-k tensor with dimensions $d^{m_1}, \dots, d^{m_k}$. For any polynomial $P\left(X^{(1)}, \dots, X^{(k)}\right)$ that is homogeneous with degree exactly m_i in the set of variables $X^{(i)}$ for all i, there is a unique tensor T such that

- *The entries of T are real numbers*
- $\langle T, A\rangle = P\left(X^{(1)}, \dots, X^{(k)}\right)$
- *The tensorization of any 1-dimensional slice of T along the i^{th} axis into a $\underbrace{d \times \cdots \times d}_{m_i}$ tensor is symmetric.*

Proof Note that each entry of T may be indexed by the corresponding monomial of A. The symmetry property implies that the entries of T that are indexed by the same monomial must be the same. Thus, the unique tensor T is constructed by taking each monomial of P and dividing its coefficient evenly among all of the entries of T that are indexed by that monomial. ■

In light of Claim 8 we may make the following definition:

Definition 33 *Let k be a positive integer and consider k distinct sets of d formal variables, say $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)}), \dots, X^{(k)} = (X_1^{(k)}, \dots, X_d^{(k)})$. For a polynomial $P\left(X^{(1)}, \dots, X^{(k)}\right)$ that is homogeneous with degree $m_1, \dots, m_k$ in the sets of variables $X^{(1)}, \dots, X^{(k)}$ respectively, let $T_{\mathsf{sym}}(P)$ be the (unique) tensor constructed in Claim 8. We call $T_{\mathsf{sym}}(P)$ the symmetric tensorization of P.*

We will need a few basic properties relating polynomials and their symmetric tensorizations. We are mostly interested in the case when there are two sets of variables (i.e. $k = 2$). In this case, the symmetric tensorizations will simply be matrices. The first property is immediate from the definition.

Claim 9 *Let k be a positive integer and consider k distinct sets of d formal variables, say $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)}), \dots, X^{(k)} = (X_1^{(k)}, \dots, X_d^{(k)})$. Let $P_1, \dots, P_k$ be homogeneous polynomials in d variables. Then*

$$T_{\mathsf{sym}}\left(P_1(X^{(1)}) \cdots P_k(X^{(k)})\right) = v(P_1(X)) \otimes \cdots \otimes v(P_k(X)) .$$

Proof Let the degrees of $P_1, \dots, P_k$ be $m_1, \dots, m_k$ respectively. Let

$$A = \mathsf{flat}\left(\left(X^{(1)}\right)^{\otimes m_1}\right) \otimes \cdots \otimes \mathsf{flat}\left(\left(X^{(k)}\right)^{\otimes m_k}\right) .$$

Note that

$$\langle A, v(P_1(X)) \otimes \cdots \otimes v(P_k(X))\rangle = P_1(X^{(1)}) \cdots P_k(X^{(k)}) .$$

Also note that each of the one dimensional slices of $v(P_1(X)) \otimes \cdots \otimes v(P_k(X))$ is symmetric when put into a $d \times \cdots \times d$ tensor. Thus, Claim 8 implies that $v(P_1(X)) \otimes \cdots \otimes v(P_k(X))$ is exactly the symmetric tensorization of $P_1(X^{(1)}) \cdots P_k(X^{(k)})$. ■

Claim 10 *Consider two sets of d formal variables, say $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)})$ and $X^{(2)} = (X_1^{(2)}, \dots, X_d^{(2)})$. Let $P(X^{(1)}, X^{(2)}), Q(X^{(1)}, X^{(2)})$ be polynomials that are homogeneous in each of the sets of variables. Then*

$$\left\|T_{sym}\left(PQ\right)\right\|_{op} \leq \left\|T_{sym}(P)\right\|_{op} \left\|T_{sym}(Q)\right\|_{op} .$$

Proof Assume that P has degrees m_1, m_2 and Q has degrees n_1, n_2 in $X^{(1)}, X^{(2)}$ respectively. Let

$$A = \mathsf{flat}\left(\left(X^{(1)}\right)^{\otimes m_1 + n_1}\right) \otimes \mathsf{flat}\left(\left(X^{(2)}\right)^{\otimes m_2 + n_2}\right) .$$

Note that

$$\langle A, T_{\mathsf{sym}}(P) \otimes T_{\mathsf{sym}}(Q) \rangle = P(X^{(1)}, X^{(2)}) Q(X^{(1)}, X^{(2)})$$

where $T_{\mathsf{sym}}(P) \otimes T_{\mathsf{sym}}(Q)$ is the Kronecker product of the two matrices. Thus, $T_{\mathsf{sym}}\left(PQ\right)$ can be written as an average of tensors that are equivalent to $T_{\mathsf{sym}}(P) \otimes T_{\mathsf{sym}}(Q)$ up to permutations of the rows and columns. Thus, by the triangle inequality

$$\left\|T_{\mathsf{sym}}\left(PQ\right)\right\|_{\mathsf{op}} \leq \left\|T_{\mathsf{sym}}(P) \otimes T_{\mathsf{sym}}(Q)\right\|_{\mathsf{op}} = \left\|T_{\mathsf{sym}}(P)\right\|_{\mathsf{op}} \left\|T_{\mathsf{sym}}(Q)\right\|_{\mathsf{op}} .$$

■

Appendix D. Hermite Polynomials, Generating Functions and Differential Operators

In this section, we introduce the Hermite moment polynomials and their associated generating functions. We then introduce several tools for manipulating generating functions using differential operators that will be crucial later on. While some of these tools were introduced in Liu and Moitra (2021), we introduce many additional tools in this paper as we will need more precise characterizations and bounds on various quantities.

D.1. Hermite Polynomials and their Generating Functions

Here we develop some basic machinery for working with Hermite polynomials and their generating functions. The first set of definitions and results mirrors the work in Liu and Moitra (2021). We begin with a standard definition.

Definition 34 *Let $\mathcal{H}_m(x)$ be the univariate Hermite polynomials $\mathcal{H}_0 = 1, \mathcal{H}_1 = x, \mathcal{H}_2 = x^2 - 1 \cdots$ defined by the recurrence*

$$\mathcal{H}_m(x) = x\mathcal{H}_{m-1}(x) - (m-1)\mathcal{H}_{m-2}(x)$$

Note that in $\mathcal{H}_m(x)$, the degree of each nonzero monomial has the same parity as m. In light of this, we can write the following:

Definition 35 *Let $\mathcal{H}_m(x, y^2)$ be the homogenized Hermite polynomials e.g. $\mathcal{H}_2(x, y^2) = x^2 - y^2, \mathcal{H}_3(x, y^2) = x^3 - 3xy^2$.*

It will be important to note the following fact:

Claim 11 *We have*

$$e^{xz-\frac{1}{2}y^2z^2} = \sum_{m=0}^{\infty} \frac{1}{m!}\mathcal{H}_m(x, y^2)z^m$$

where the RHS is viewed as a formal power series in z whose coefficients are polynomials in x, y.

Now we define a multivariate version of the Hermite polynomials.

Definition 36 (Multivariate Hermite Polynomials) *Let $H_m(X, z)$ be a formal polynomial in variables $X = (X_1, \dots, X_d)$ whose coefficients are polynomials in d variables $z_1, \dots, z_d$ that is given by*

$$H_m(X, z) = \mathcal{H}_m(z_1X_1 + \cdots + z_dX_d, X_1^2 + \cdots + X_d^2)$$

We call $H_m(X, z)$ the multivariate Hermite polynomials. Note that H_m is homogeneous of degree m as a polynomial in $X_1, \dots, X_d$

Definition 37 (Hermite Moment Polynomials) *For a distribution D on $\mathbb{R}^d$, we let*

$$h_{m,D}(X) = \mathbb{E}_{(z_1,\dots,z_d)\sim D}[H_m(X, z)]$$

where we take the expectation of H_m over $(z_1, \dots, z_d)$ drawn from D. Note that $h_{m,D}(X)$ is a polynomial in d variables $(X_1, \dots, X_d)$. We will omit the D in the subscript when it is clear from context. We refer to $h_{m,D}(X)$ as the Hermite moment polynomials of D.

We can extend the above definition to any function $f : \mathbb{R}^d \to \mathbb{R}$ (that is not necessarily a distribution).

Definition 38 *For any function $f : \mathbb{R}^d \to \mathbb{R}$, we define*

$$h_{m,f}(X) = \int_{\mathbb{R}^d} f(z)H_m(X, z)dz\,.$$

Remark 39 *Note that there is no ambiguity because this definition agrees with the above when f is a distribution. The extended definition will mostly be used for working with MPG functions that may not be normalized and may take on negative values.*

The first important observation is that the Hermite moment polynomials for a Gaussian can be written in a simple closed form via generating functions.

Claim 12 *Let $D = N(\mu, I + \Sigma)$. Then*

$$\sum_{m=0}^{\infty} \frac{1}{m!} \cdot h_{m,D}(X)y^m = e^{\mu(X)y+\frac{1}{2}\Sigma(X)y^2}$$

where both sides are viewed as formal power series in y whose coefficients are polynomials in X.

Proof Using Claim 11, the LHS may be rewritten as

$$\mathbb{E}_{z\sim D}\left[\sum_{m=0}^{\infty}\frac{1}{m!}\cdot H_m(X,z)y^m\right]=\mathbb{E}_{z\sim D}\left[e^{(z_1X_1+\cdots+z_dX_d)y-\frac{1}{2}(X_1^2+\cdots+X_d^2)y^2}\right]$$
$$=C\int\exp\left(-\frac{1}{2}(z-\mu)^T(I+\Sigma)^{-1}(z-\mu)+z^TXy-\frac{1}{2}X^TXy^2\right)dz$$
$$=C\int\exp\left(-\frac{1}{2}(z-\mu-(I+\Sigma)Xy)^T(I+\Sigma)^{-1}(z-\mu-(I+\Sigma)Xy)+\mu^TXy+\frac{1}{2}X^T\Sigma Xy^2\right)dz$$
$$=\exp\left(\mu(X)y+\frac{1}{2}\Sigma(X)y^2\right).$$

where in the above, C is the normalization constant for a normal distribution with covariance $I+\Sigma$. Note that for the last step, we used the fact that

$$\int\exp\left(\frac{1}{2}(z-\mu)^T(I+\Sigma)^{-1}(z-\mu)\right)dz$$
$$=\int\exp\left(\frac{1}{2}(z-\mu-(I+\Sigma)Xy)^T(I+\Sigma)^{-1}(z-\mu-(I+\Sigma)Xy)\right)dz.$$

■

By slightly modifying the proof of Claim 12, we can prove a more general result when we have a function given by a polynomial Gaussian.

Claim 13 *Let $f(x):\mathbb{R}^d\to\mathbb{R}$ be given by $f(x)=Q(x)G(x)$ where $G=N(\mu,I+\Sigma)$ is a Gaussian and Q is a polynomial of degree c. Then*

$$\sum_{m=0}^{\infty}\frac{1}{m!}\cdot h_{m,f}(X)y^m=P(Xy)e^{\mu(X)y+\frac{1}{2}\Sigma(X)y^2}$$

where P is a polynomial in d variables of degree at most c and Xy denotes the d-tuple of formal variables $(X_1y,\dots,X_dy)$.

Proof Using Claim 11, the LHS may be rewritten as

$$\int_{\mathbb{R}^d}f(z)\left[\sum_{m=0}^{\infty}\frac{1}{m!}\cdot H_m(X,z)y^m\right]dz=\int_{\mathbb{R}^d}f(z)\left[e^{(z_1X_1+\cdots+z_dX_d)y-\frac{1}{2}(X_1^2+\cdots+X_d^2)y^2}\right]dz$$
$$=C\int Q(z)\exp\left(-\frac{1}{2}(z-\mu)^T(I+\Sigma)^{-1}(z-\mu)+z^TXy-\frac{1}{2}X^TXy^2\right)dz$$
$$=C\int Q(z)\exp\left(-\frac{1}{2}(z-\mu-(I+\Sigma)Xy)^T(I+\Sigma)^{-1}(z-\mu-(I+\Sigma)Xy)+\mu^TXy+\frac{1}{2}X^T\Sigma Xy^2\right)dz$$
$$=Ce^{\mu(X)y+\frac{1}{2}\Sigma(X)y^2}\int Q(z)\exp\left(-\frac{1}{2}(z-\mu-(I+\Sigma)Xy)^T(I+\Sigma)^{-1}(z-\mu-(I+\Sigma)Xy)\right)dz.$$

where in the above, C is a constant. Now we can make the change of variables $z = (I+\Sigma)^{1/2}z' + \mu + (I+\Sigma)Xy$ and deduce that

$$\begin{aligned}\int Q(z)\exp\left(-\frac{1}{2}(z-\mu-(I+\Sigma)Xy)^T(I+\Sigma)^{-1}(z-\mu-(I+\Sigma)Xy)\right)dz \\ = \det(I+\Sigma)^{1/2}\int Q\left((I+\Sigma)^{1/2}z+\mu+(I+\Sigma)Xy\right)\exp\left(-\frac{1}{2}\|z\|^2\right)dz \\ = P(\mu+(I+\Sigma)Xy)\end{aligned}$$

for some polynomial P of degree at most c. Putting everything together gives us the desired result. ■

We now have a few simple consequences of the above.

Corollary 40 *Let $\mathcal{M} = w_1G_1 + \dots w_kG_k$ be a mixture of Gaussians where $G_j = N(\mu_j, I+\Sigma_j)$. Then*

$$\sum_{m=0}^{\infty}\frac{1}{m!}\cdot h_{m,\mathcal{M}}(X)y^m = w_1e^{\mu_1(X)y+\frac{1}{2}\Sigma_1(X)y^2} + \dots + w_ke^{\mu_k(X)y+\frac{1}{2}\Sigma_k(X)y^2}$$

Corollary 41 *Let $f(x) = Q_1(x)G_1(x) + \dots + Q_k(x)G_k(x)$ be a degree-c MPG function in $\mathbb{R}^d$ where $G_j = N(\mu_j, I+\Sigma_j)$. Then there are polynomials $P_1(X), \dots, P_k(X)$ of degree at most c in formal variables $X = (X_1, \dots, X_d)$ such that*

$$\sum_{m=0}^{\infty}\frac{1}{m!}h_{m,f}(X)y^m = P_1(Xy)e^{\mu_1(X)y+\frac{1}{2}\Sigma_1(X)y^2} + \dots + P_k(Xy)e^{\mu_k(X)y+\frac{1}{2}\Sigma_k(X)y^2}$$

where Xy denotes the d-tuple of formal variables $(X_1y, \dots, X_dy)$.

D.2. The Adjusted Characteristic Function and its Properties

The characteristic function is a well-known concept in probability. Here, we use a modified notion of an adjusted characteristic function. One of the key components in our paper is Theorem 50, which relates the L^1 distance between MPG functions (note this is equivalent to TV distance for mixtures of Gaussians and MPG distributions) to the coefficient-norm distance between their Hermite moment polynomials. The adjusted characteristic function will play a key role in proving Theorem 50 because its inverse map gives us a way to map from a generating function for Hermite moment polynomials back to a distribution.

Definition 42 *For a function f on $\mathbb{R}^d$, we define its adjusted characteristic function $\widetilde{f} : \mathbb{R}^d \to \mathbb{C}$ as*

$$\widetilde{f}(X) = \int_{\mathbb{R}^d} f(z)\left[e^{iz\cdot X+\frac{1}{2}\|X\|^2}\right]dz$$

where $i = \sqrt{-1}$.

Note that for distributions, the adjusted characteristic function is the characteristic function multiplied by $\frac{1}{2}\|X\|^2$. Now we define the inverse map.

Definition 43 *For a function $g : \mathbb{R}^d \to \mathbb{C}$, we define χg, a function from $\mathbb{R}^d$ to $\mathbb{C}$, as follows:*

$$\chi g(t) = \frac{1}{(2\pi)^d} \int g(X) e^{-\frac{1}{2}\|X\|^2 - it\cdot X} dX$$

where in the integral above, X ranges over all of $\mathbb{R}^d$.

It is straight-forward to verify that the transformation defined above indeed inverts the adjusted characteristic function.

Fact 1 *For a function f on $\mathbb{R}^d$,*

$$\chi \widetilde{f} = f\,.$$

A key property of the adjusted characteristic function is that its output is equivalent to plugging in $y = i$ into the generating function for its Hermite moment polynomials.

Claim 14 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. Then*

$$\widetilde{f}(X) = \sum_{m=0}^{\infty} \frac{i^m}{m!} h_{m,f}(X)\,.$$

Proof Note that by Claim 11,

$$\sum_{m=0}^{\infty} \frac{i^m}{m!} h_{m,f}(X) = \int_{\mathbb{R}^d} f(z) e^{iz\cdot X + \frac{1}{2}\|X\|^2} dz = \widetilde{f}(X)\,,$$

as desired. ■

As a consequence of the above and Corollary 41, we have

Corollary 44 *If we have a degree-c MPG function $f = Q_1(x)G_1 + \cdots + Q_k(x)G_k$ then*

$$\widetilde{f}(X) = P_1(iX) e^{i\mu_1(X) - \frac{1}{2}\Sigma_1(X)} + \cdots + P_k(iX) e^{i\mu_k(X) - \frac{1}{2}\Sigma_k(X)}$$

for some polynomials $P_1, \dots, P_k$ of degree at most c with real coefficients.

In light of the above, we know that the adjusted characteristic function maps a function to a generating function for its Hermite moment polynomials. Recall that our goal is to prove that small distance between Hermite moment polynomials implies small TV distance. This means that we need to understand the L^1 norm of the inverse adjusted characteristic function. In the remainder of this subsection, we prove some basic quantitative bounds on the inverse adjusted characteristic function that will be used later on.

D.2.1. COMPUTATIONS IN 1D

The following two identities follow from direct computation.

Claim 15 *For a real number $t \in \mathbb{R}$,*

$$\int_{-\infty}^{\infty} x^m e^{-\frac{1}{2}x^2} e^{-itx} dx = (-i)^m \sqrt{2\pi} e^{-t^2/2} \mathcal{H}_m(t)$$

where recall $\mathcal{H}_m(x)$ is the univariate Hermite polynomial.

Fact 2

$$\int_{-\infty}^{\infty} \mathcal{H}_{m_1}(x) \mathcal{H}_{m_2}(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = 1_{m_1 = m_2} (m_1)!$$

D.2.2. BOUNDS ON THE INVERSE ADJUSTED CHARACTERISTIC FUNCTION

The next result gives us several important properties for certain inverse adjusted characteristic functions corresponding to polynomial Gaussians.

Claim 16 *Let $p(X)$ be a polynomial with real coefficients in d variables $X_1, \dots, X_d$ that is homogeneous of degree m. Consider a Gaussian in $\mathbb{R}^d$, $G = N(\mu, I + \Sigma)$. For $X \in \mathbb{R}^d$ let*

$$g(X) = i^m p(X) e^{i\mu(X) - \frac{1}{2}\Sigma(X)}\,.$$

Then

$$\chi g(t) = q(t)G(t)$$

for some polynomial q of degree at most m with real coefficients. Furthermore,

$$\mathbb{E}_{t \sim G}\left[q(t)^2\right] \leq m!(\left\|(I+\Sigma)^{-1}\right\|_{op})^m \left\|v(p)\right\|^2$$

where recall $v(p)$, defined in Definition 30, is the vectorization of the coefficients of p.

Proof

We have

$$\chi g(t) = \frac{1}{(2\pi)^d} \int i^m p(X) e^{-\frac{1}{2}X^T(I+\Sigma)X - i(t-\mu)\cdot X} dX$$

Substituting $X \to (I+\Sigma)^{-1/2}Y$ for $Y = (Y_1, \dots, Y_d)$ in the above we get

$$\chi g(t) = \frac{1}{(2\pi)^d \det(I+\Sigma)^{1/2}} \int i^m p((I+\Sigma)^{-1/2}Y) e^{-\frac{1}{2}\|Y\|^2 - i(I+\Sigma)^{-1/2}(t-\mu)\cdot Y} dY\,.$$

To compute the above integral, note that $p((I+\Sigma)^{-1/2}Y)$ is a polynomial in $Y_1, \dots, Y_d$ that is homogeneous of degree m. Let $h(Y) = p((I+\Sigma)^{-1/2}Y)$. Let $s = (I+\Sigma)^{-1/2}(t-\mu)$ and let its coordinates be $s_1, \dots, s_d$. We now separate h into monomials and consider one monomial at a time. Consider a monomial say $Y_1^{a_1}Y_2^{a_2}\cdots Y_d^{a_d}$ for some integers $a_1, \dots, a_d$. Note that the term inside the exponential can be factored coordinate-wise. Thus, we can apply Claim 15 to compute the integral as follows:

$$\begin{aligned}\int i^m Y_1^{a_1} \cdots Y_d^{a_d} e^{-\frac{1}{2}\|Y\|^2 - is\cdot Y} dY &= i^m \prod_{j=1}^d \int_{-\infty}^{\infty} x^{a_j} e^{-\frac{1}{2}x^2} e^{-is_j x} dx \\ &= (2\pi)^{d/2} e^{-\frac{1}{2}(s_1^2 + \cdots + s_d^2)} \prod_{j=1}^d \mathcal{H}_{a_j}(s_j)\,.\end{aligned}$$

We denote the coefficients of h by $c_{a_1,\dots,a_d}$. Combining the above over all monomials, we get

$$\begin{aligned}\chi g(t) &= \frac{1}{(2\pi)^{d/2}\det(I+\Sigma)^{1/2}} e^{-\frac{1}{2}\|s\|^2} \sum_{a_1,\dots,a_d} c_{a_1,\dots,a_d} \prod_{j=1}^d \mathcal{H}_{a_j}(s_j) \\ &= \left(\sum_{a_1,\dots,a_d} c_{a_1,\dots,a_d} \prod_{j=1}^d \mathcal{H}_{a_j}(s_j)\right) N(\mu, I+\Sigma)(t)\,.\end{aligned}$$

From the above we deduce

$$q(t) = \sum_{a_1,\dots,a_d} c_{a_1,\dots,a_d} \prod_{j=1}^{d} \mathcal{H}_{a_j}(s_j)$$

and it is clear that q is a polynomial of degree at most m in t (since s is a linear function of t). It remains to bound $\mathbb{E}_{t\sim G}\left[q(t)^2\right]$. Note that

$$\begin{aligned}
\mathbb{E}_{t\sim G}\left[q(t)^2\right] &= \int_{\mathbb{R}^d} \left(\sum_{a_1,\dots,a_d} c_{a_1,\dots,a_d} \prod_{j=1}^{d} \mathcal{H}_{a_j}(s_j) \right)^2 N(\mu, I+\Sigma)(t)dt \\
&= \int_{\mathbb{R}^d} \left(\sum_{a_1,\dots,a_d} c_{a_1,\dots,a_d} \prod_{j=1}^{d} \mathcal{H}_{a_j}(s_j) \right)^2 N(0, I)(s)ds \\
&= \sum_{a_1,\dots,a_d} \sum_{a'_1,\dots,a'_d} \int_{\mathbb{R}^d} c_{a_1,\dots,a_d} c_{a'_1,\dots,a'_d} \prod_{j=1}^{d} \mathcal{H}_{a_j}(s_j) H_{a'_j}(s_j) \frac{e^{-\|s\|^2/2}}{(2\pi)^{d/2}} ds\,.
\end{aligned}$$

However, since the integral factorizes over the different coordinates of s, by Fact 2, the integral evaluates to 0 unless $a_1,\dots,a_d = a'_1,\dots,a'_d$ and overall, we get

$$\begin{aligned}
\int_{\mathbb{R}^d} \left(\sum_{a_1,\dots,a_d} c_{a_1,\dots,a_d} \prod_{j=1}^{d} \mathcal{H}_{a_j}(s_j) \right)^2 \frac{e^{-\|s\|^2/2}}{(2\pi)^{d/2}} ds &= \sum_{a_1,\dots,a_d} c^2_{a_1,\dots,a_d} \prod_{j=1}^{d} \int_{-\infty}^{\infty} \mathcal{H}_{a_j}(x)^2 \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \\
= \sum_{a_1,\dots,a_d} c^2_{a_1,\dots,a_d} \prod_{j=1}^{d} a_j! = m! \sum_{a_1,\dots,a_d} \binom{m}{a_1,\dots,a_d} &\left(\frac{c_{a_1,\dots,a_d}}{\binom{m}{a_1,\dots,a_d}} \right)^2 = m! \left\| v(h) \right\|^2 .
\end{aligned}$$

Finally, by Claim 7, $\left\| v(h) \right\|^2 \leq (\left\|(I+\Sigma)^{-1}\right\|_{\mathsf{op}})^m \left\| v(p) \right\|^2$ so combining with the previous inequality, we deduce

$$\mathbb{E}_{t\sim G}\left[q(t)^2\right] \leq m!(\left\|(I+\Sigma)^{-1}\right\|_{\mathsf{op}})^m \left\| v(p) \right\|^2 .$$

■

As a consequence of the previous claim, we have

Corollary 45 *Let $G_1 = N(\mu_1, I+\Sigma_1), \dots, G_k = N(\mu_k, I+\Sigma_k)$ be Gaussians. Let c be a constant. There is a one-to-one correspondence between polynomials $Q_1,\dots,Q_k$ of degree at most c with real coefficients and polynomials $P_1,\dots,P_k$ of degree at most c with real coefficients given by the following map:*

The adjusted characteristic function of $f(x) = Q_1(x)G_1 + \dots + Q_k(x)G_k$ is

$$\widetilde{f}(X) = P_1(iX)e^{i\mu_1(X) - \frac{1}{2}\Sigma_1(X)} + \dots + P_k(iX)e^{i\mu_k(X) - \frac{1}{2}\Sigma_k(X)} .$$

Proof This follows immediately from Corollary 44 and Claim 16. ■

Claim 16 also implies a bound on the L^1 norm of the inverse adjusted characteristic function of a polynomial in terms of the size of its coefficients.

Claim 17 *Let $p(X)$ be a polynomial in d variables $X_1, \dots, X_d$ that is homogeneous of degree m with real coefficients. Then*

$$\|\chi(i^m p(X))\|_1 \leq \sqrt{m!}\, \|v(p)\|_2 \, .$$

Proof By Claim 16 (with $\mu = 0, \Sigma = 0$),

$$\chi(i^m p)(t) = q(t) N(0, I)(t)$$

where q is a polynomial of degree at most m such that $\mathbb{E}_{t \sim N(0,I)}[q(t)^2] \leq m! \, \|v(p)\|^2$. Now

$$\|\chi(i^m p)\|_1 = \int_{\mathbb{R}^d} |q(t)| N(0,I)(t) dt \leq \sqrt{\int_{\mathbb{R}^d} |q(t)|^2 N(0,I)(t) dt} \leq \sqrt{m!}\, \|v(p)\|$$

where we used Cauchy Schwarz in the first inequality. ■

D.3. Generating Function Terminology

Here we introduce some general terminology for working with generating functions related to mixtures of Gaussians and mixtures of polynomial Gaussians. In light of Corollaries 40 and 41, it will be useful to translate between generating functions consisting of sums of exponentials and their expansions as formal power series in y whose coefficients are polynomials in X e.g.

$$f(X, y) = \sum_{j=0}^{\infty} \frac{f_j(X)}{j!} y^j \, .$$

For such an expression, we use the following terminology.

Definition 46 *Given a formal power series in y, say*

$$f(X, y) = \sum_{j=0}^{\infty} \frac{f_j(X)}{j!} y^j \, ,$$

where the coefficients $f_0(X), f_1(X), \dots$ are polynomials in formal variables $X = (X_1, \dots, X_d)$ and have real coefficients, we call the polynomials $f_0(X), f_1(X), \dots$ the primary terms of f.

We also introduce terminology for dealing with mixtures of polynomial Gaussians.

Definition 47 *Given Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$ in $\mathbb{R}^d$, and polynomials $P_1(X), \dots, P_k(X)$ in formal variables $X = (X_1, \dots, X_d)$ with real coefficients, we say that the generating function of the polynomial combination is*

$$f(X, y) = P_1(Xy) e^{\mu_1(X)y + \frac{1}{2}\Sigma_1(X)y^2} + \dots + P_k(Xy) e^{\mu_k(X)y + \frac{1}{2}\Sigma_k(X)y^2}$$

where we view f as a function of y with indeterminates X. Recall that Xy denotes the d-tuple of variables $(X_1 y, \dots, X_d y)$.

Remark 48 *Note that by Corollary 40, if we have a mixture of Gaussians $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$, then the generating function of the polynomial combination $P_1(X) = w_1, \dots, P_k(X) = w_k$ is*

$$f(X, y) = w_1 e^{\mu_1(X)y + \frac{1}{2}\Sigma_1(X)y^2} + \cdots + w_k e^{\mu_k(X)y + \frac{1}{2}\Sigma_k(X)y^2} = \sum_{m=0}^{\infty} \frac{1}{m!} h_{m,\mathcal{M}}(X) y^m \, .$$

In other words, the primary terms in the formal power series expansion of $f(y)$ are exactly the Hermite moment polynomials of the mixture. More generally, Corollary 41 and Corollary 45 imply that for an MPG function $\mathcal{M} = Q_1(x)G_1 + \cdots + Q_k(x)G_k$, there are corresponding polynomials $P_1, \dots, P_k$, such that the generating function of the polynomial combination is

$$f(X, y) = P_1(Xy) e^{\mu_1(X)y + \frac{1}{2}\Sigma_1(X)y^2} + \cdots + P_k(Xy) e^{\mu_k(X)y + \frac{1}{2}\Sigma_k(X)y^2} = \sum_{m=0}^{\infty} \frac{1}{m!} h_{m,\mathcal{M}}(X) y^m \, .$$

In the exposition, we will use the following (informal) terminology. Note that for a Gaussian $N(\mu, I + \Sigma)$, we often associate it with a generating function of the form $e^{\mu(X)y + \frac{1}{2}\Sigma(X)y^2}$.

- When we work with the actual pdfs of Gaussians and write e.g. expressions of the form $f = Q_1(x)G_1 + \cdots + Q_k(x)G_k$, we say that we are working in **distribution space**
- When we work with expressions of the form $e^{\mu(X)y + \frac{1}{2}\Sigma(X)y^2}$ and e.g. expand them as power series containing Hermite moment polynomials, we say that we are working in **generating function space**

D.4. Differential Operators and their Compositions

Before moving on to the main proofs, we need to introduce one more piece of machinery: differential operators. Later on, differential operators will play a crucial role in allowing us to manipulate generating functions and derive useful identities. In this section, we present a few basic results that will be used throughout the paper.

We will frequently work with operators given by

$$\mathcal{D} = \partial - (a(X) + b(X)y)$$

where the partial derivative is taken with respect to y, $a(X)$ is a (homogeneous) linear function and $b(X)$ is a (homogeneous) quadratic function. Note that if applied to a formal power series

$$f(y) = \sum_{j=0}^{\infty} \frac{Q_j(X)}{j!} y^j \, ,$$

the terms of the resulting power series are

$$\mathcal{D}(f(y)) = \sum_{j=0}^{\infty} \frac{R_j(X)}{j!} y^j$$

where $R_j(X) = Q_{j+1}(X) - a(X)Q_j(X) - jb(X)Q_{j-1}(X)$. In particular, if for all j, Q_j is homogeneous of degree j in X, then the primary terms $R_j(X)$ are homogeneous and of degree $j + 1$.

It will be important to understand compositions of differential operators. We now prove several basic properties that will be used later on. The first claim allows us to rewrite a composition of differential operators as a higher order differential operator with polynomials as coefficients.

Claim 18 *Consider a composition of differential operators*

$$\mathcal{D} = (\partial - (a_k(X) + b_k(X)y)) \cdots (\partial - (a_1(X) + b_1(X)y))$$

where each a_j is linear and each b_j is quadratic. Then $\mathcal{D}$ can be rewritten in the form

$$\partial^k + R_{k-1}(X,y)\partial^{k-1} + \cdots + R_0(X,y)$$

where

- *Each R_j is a polynomial of degree at most $k-j$ in y*

$$R_j(X,y) = R_{j,k-j}(X)y^{k-j} + \cdots + R_{j,0}(X)$$

- *Each of the polynomials $R_{j,l}$ is homogeneous in X with degree $k-j+l$, and is $(O_k(1),k)$-simple with respect to $\{a_1(X), b_1(X), \dots, a_k(X), b_k(X)\}$.*

Proof We will use induction on k. The base case is clear. Now, assume that we have written the operator

$$\mathcal{D}_{k-1} = (\partial - (a_{k-1}(X) + b_{k-1}(X)y)) \cdots (\partial - (a_1(X) + b_1(X)y))$$

in the desired form

$$\mathcal{D}_{k-1} = \partial^{k-1} + R_{k-2}(X,y)\partial^{k-2} + \cdots + R_0(X,y)\,.$$

When we apply the last differential operator, we get

$$\begin{aligned}(\partial - (a_k(X) + b_k(X)y))\mathcal{D}_{k-1} = {} & \partial(\partial^{k-1} + R_{k-2}(X,y)\partial^{k-2} + \cdots + R_0(X,y)) \\ & - (a_k(X) + b_k(X)y)(\partial^{k-1} + R_{k-2}(X,y)\partial^{k-2} + \cdots + R_0(X,y))\,.\end{aligned}$$

It is clear that the second term can be written in the desired form. To deal with the first term, we can simply use the product rule i.e.

$$\partial(R_j(X,y)\partial^j) = R_j(X,y)\partial^{j+1} + \partial(R_j(X,y))\partial^j$$

to write the entire differential operator in the desired form.

■

The next claim implies that applying a differential operator of the form $\partial - (a(X) + b(X)y)$ cannot annihilate (or nearly annihilate) a polynomial unless $a(X), b(X)$ are both close to 0.

Claim 19 *Consider a polynomial $P(X,y)$ of degree k in y*

$$P(X,y) = P_0(X) + P_1(X)y + \cdots + P_k(X)y^k$$

where the coefficients $P_j(X)$ are polynomials in X that are homogeneous of degree $k'+j$ for some constant k'. Let

$$R(X,y) = -(a(X)+b(X)y)P(X,y) + \partial(P(X,y))$$

where $a(X)$, $b(X)$ are a (homogeneous) linear and quadratic respectively and the partial derivative is taken with respect to y. Assume

$$\delta \leq \|v(a(X))\| + \|v(b(X))\| \leq \delta^{-1}\,.$$

Then

$$\|v_y(R(X,y))\| \geq (0.1\delta)^{O_{k,k'}(1)} \|v_y(P(X,y))\|\ .$$

Proof We will induct on k, the degree of P. The case when $k=0$ follows from Claim 6. Now assume that there is a constant $c_{k-1,k'}$ for which the desired statement holds for all polynomials P of degree at most $k-1$ in y. Note that the coefficients of y^{k+1}, y^k in R are

$$\begin{aligned} R_{k+1}(X) &= -b(X)P_k(X) \\ R_k(X) &= -b(X)P_{k-1}(X) - a(X)P_k(X) \end{aligned}$$

Thus by Claim 5 and Claim 6,

$$\begin{aligned} \|v(R_{k+1}(X))\| &\geq \Omega_{k,k'}(1)\|v(b(X))\|\|v(P_k(X))\| \\ \|v(R_k(X))\| &\geq \Omega_{k,k'}(1)\|v(a(X))\|\|v(P_k(X))\| - O_{k,k'}(1)\|v(b(X))\|\|v(P_{k-1}(X))\| \end{aligned}$$

If for some parameter ϵ,

$$\|v(P_k(X))\| \geq \epsilon \|v_y(P(X,y)\|$$

then for some sufficiently large constant K depending only on k, k',

$$\begin{aligned} \|v_y(R(X,y))\| &\geq \frac{1}{2}\|v(R_{k+1}(X))\| + \frac{\epsilon}{K}\|v(R_k(X))\| \\ &\geq \|v(b(X))\|\left(\Omega_{k,k'}(1)\|v(P_k(X))\| - \frac{\epsilon}{K}O_{k,k'}(1)\|v(P_{k-1}(X))\|\right) \\ &\quad + \Omega_{k,k'}(1)\frac{\epsilon}{K}\|v(a(X))\|\|v(P_k(X))\| \\ &\geq \Omega_{k,k'}(1)\frac{\epsilon^2}{K}(\|v(a(X))\| + \|v(b(X))\|)\|v_y(P(X,y))\| \\ &\geq \Omega_{k,k'}(1)\epsilon^2\delta\|v_y(P(X,y))\|\ . \end{aligned}$$

On the other hand, if ϵ is sufficiently small and

$$\|v(P_k(X))\| \leq \epsilon \|v_y(P(X,y)\|\ ,$$

we may use the induction hypothesis on the polynomial

$$P'(X,y) = P_0(X) + P_1(X)y + \cdots + P_{k-1}(X)y^{k-1}\,.$$

Note

$$R(X,y) = -(a(X)+b(X)y)P'(X,y) + \partial(P'(X,y)) - (a(X)+b(X)y)P_k(X)y^k + kP_k(X)y^{k-1}$$

so by the induction hypothesis

$$\begin{aligned}\|v_y(R(X,y))\| &\geq (0.1\delta)^{c_{k-1,k'}} \left\|v_y(P'(X,y))\right\| - O_{k,k'}(1)(1 + \|v(a(X))\| + \|v(b(X))\|) \|v(P_k(X))\| \\ &\geq \left((0.1\delta)^{c_{k-1,k'}}(1-\epsilon) - O_{k,k'}(1)\epsilon\delta^{-1}\right) \|v_y(P(X,y))\| \ .\end{aligned}$$

Choosing $\epsilon = (0.1\delta)^{c_{k,k'}}$ for $c_{k,k'}$ sufficiently large in terms of $k, k', c_{k-1,k'}$, in both cases we get

$$\|v_y(R(X,y))\| \geq (0.1\delta)^{O_{k,k'}(1)} \|v_y(P(X,y))\| \ ,$$

completing the proof. ■

By repeatedly applying the previous claim, we can lower bound the order-0 term when a composition of differential operators $(\partial - (a_k(X) + b_k(X)y)) \cdots (\partial - (a_1(X) + b_1(X)y))$ is expanded into an order-k differential operator with polynomial coefficients (recall Claim 18).

Claim 20 *Consider a composition of differential operators*

$$\mathcal{D} = (\partial - (a_k(X) + b_k(X)y)) \cdots (\partial - (a_1(X) + b_1(X)y))$$

where each a_j is linear and each b_j is quadratic. Assume that $\mathcal{D}$ is rewritten in the form

$$\mathcal{D} = \partial^k + R_{k-1}(X,y)\partial^{k-1} + \cdots + R_0(X,y) \, .$$

Also assume that for some constant δ,

$$\delta < \|v(a_j(X))\| + \|v(b_j(X))\| < \delta^{-1}$$

for all j. There exists a (sufficiently large) constant C_k depending only on k such that

$$\|v_y(R_0(X,y))\| \geq (0.1\delta)^{C_k} \, .$$

Proof We will again use induction on k. The base case is clear. Now write

$$\begin{aligned}\mathcal{D}_{k-1} &= (\partial - (a_{k-1}(X) + b_{k-1}(X)y)) \cdots (\partial - (a_1(X) + b_1(X)y)) \\ &= \partial^{k-1} + T_{k-2}(X,y)\partial^{k-2} + \cdots + T_0(X,y)\end{aligned}$$

for some polynomials $T_0, T_1, \ldots, T_{k-2}$ satisfying the conditions in Claim 18. The induction hypothesis gives us

$$\|v_y(T_0(X,y))\| \geq (0.1\delta)^{C_{k-1}} \, .$$

Now

$$R_0(X,y) = -(a_k(X) + b_k(X)y))T_0(X,y) + \partial(T_0(X,y)) \, ,$$

and applying Claim 19 completes the induction.

■

D.5. Applying Differential Operators to Generating Functions

In this section, we analyze what happens when we apply certain differential operators to specific types of generating functions. The results in this section are all from Liu and Moitra (2021) and can be verified through direct computation.

Claim 21 *Let ∂ denote differentiation with respect to y. If*

$$f(y) = P(y, X)e^{a(X)y+\frac{1}{2}b(X)y^2}$$

where P is a polynomial in y of degree k (whose coefficients are polynomials in X) then

$$(\partial - (a(X) + yb(X)))f(y) = Q(y, X)e^{a(X)y+\frac{1}{2}b(X)y^2}$$

where Q is a polynomial in y with degree exactly $k - 1$ whose leading coefficient is k times the leading coefficient of P.

Corollary 49 *Let ∂ denote the differential operator with respect to y. If*

$$f(y) = P(y, X)e^{a(X)y+\frac{1}{2}b(X)y^2}$$

where P is a polynomial in y of degree k then

$$(\partial - (a(X) + yb(X)))^{k+1}f(y) = 0.$$

Appendix E. Strong Observability: Hermite Moments to TV Distance

In this section, we prove our main observability theorem that if two degree-m MPG functions with k components are ϵ-close on their first $O_{k,m}(1)$ Hermite moment polynomials, then the functions must be $\widetilde{O}(\epsilon)$-close in L^1 distance. The theorem is stated formally below.

Theorem 50 *Let $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$ be a set of Gaussians in (α, β)-regular form. Let $Q_1(X), \dots, Q_k(X)$ be d-variate polynomials of degree at most m. Define the function $g : \mathbb{R}^d \to \mathbb{R}$ as*

$$g(x) = Q_1(x)G_1(x) + \cdots + Q_k(x)G_k(x)\,.$$

There exists a constant C depending only on k, m such that the following holds. If for all j with $0 \le j \le C$, we have

$$\|v(h_{j,g}(X))\| \le \epsilon$$

then we must have

$$\|g\|_1 \le (2 + \alpha + \beta)^C \epsilon\,.$$

Remark 51 *Note that the above gives observability because we can simply set g to be equal to the difference of two MPG-functions.*

In this section, we will use the following conventions.

- We have Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$
- $\mathcal{D}_j$ denotes the differential operator $(\partial - (\mu_j(X) + \Sigma_j(X)y))$ where the partial derivative is taken with respect to y.

E.1. Recurrence for Hermite Moment Polynomials

One of the key ingredients in the proof of Theorem 50 is writing down a recurrence for the Hermite moment polynomials of an MPG function. This is done in the following lemma.

Lemma 52 *Let $G_1 = N(\mu_1, I+\Sigma_1), \dots, G_k = N(\mu_k, I+\Sigma_k)$ be Gaussians and $Q_1(X), \dots, Q_k(X)$ be d-variate polynomials of degree at most m. Let*

$$g(x) = Q_1(x)G_1(x) + \cdots + Q_k(x)G_k(x)\,.$$

Let $\kappa = (m+1)(2^k - 1)$. Then there are d-variate polynomials $R_{j,l}(X)$ for $0 \leq j \leq \kappa$ and $0 \leq l \leq \kappa - j$ such that

- *$R_{j,l}(X)$ is homogeneous of degree $\kappa - j + l$, $R_{\kappa,0}(X) = 1$*
- *$R_{j,l}(X)$ is $(O_{k,m}(1), O_{k,m}(1))$-simple with respect to $\{\mu_1(X), \Sigma_1(X), \dots, \mu_k(X), \Sigma_k(X)\}$*
- *For all integers $a \geq \kappa$,*

$$\sum_{j=0}^{\kappa}\sum_{l=0}^{\kappa-j} \frac{h_{a-\kappa+j-l,g}(X)R_{j,l}(X)}{(a-\kappa-l)!} = 0\,,$$

where undefined terms (i.e. negative factorials in the denominator) are treated as 0.

Proof By Corollary 41, we can write

$$f(y) = \sum_{j=0}^{\infty}\frac{1}{j!}h_{j,g}(X)y^j = P_1(Xy)e^{\mu_1(X)y+\frac{1}{2}\Sigma_1(X)y^2} + \cdots + P_k(Xy)e^{\mu_k(X)y+\frac{1}{2}\Sigma_k(X)y^2}\,,$$

where $P_1, \dots, P_k$ are polynomials of degree at most m. Now consider the differential operator

$$\mathcal{D} = \mathcal{D}_k^{(m+1)2^{k-1}}\mathcal{D}_{k-1}^{(m+1)2^{k-2}}\dots\mathcal{D}_1^{m+1}\,.$$

By Claim 21, we know $\mathcal{D}(f) = 0$. However, we can expand out the formula for $\mathcal{D}$ using Claim 18 and write it in the following form

$$\mathcal{D} = \partial^\kappa + R_{\kappa-1}(X,y)\partial^{\kappa-1} + \cdots + R_1(X,y)\partial + R_0(X,y)\,.$$

Note that for each $0 \leq j \leq \kappa$, $R_j(X,y)$ is a polynomial of degree at most $\kappa - j$ in y. Furthermore, by Claim 18, it can be written in the form

$$R_j(X,y) = \sum_{l=0}^{\kappa-j} R_{j,l}(X)y^l\,,$$

where each of the polynomials $R_{j,l}$ is homogeneous of degree $\kappa - j + l$ in X and is $(O_{k,m}(1), O_{k,m}(1))$-simple with respect to $\{\mu_1(X), \Sigma_1(X) \dots, \mu_k(X), \Sigma_k(X)\}$. Furthermore, it is obvious that $R_{\kappa,0} = 1$

Using the fact that $\mathcal{D}(f) = 0$, we get that the polynomials $h_{j,g}$ in the generating function satisfy a recurrence relation of depth $O_{k,m}(1)$. For any integer a, by looking at the coefficient of $y^{a-\kappa}$ in the power series expansion of $\mathcal{D}(f)$, we deduce

$$\sum_{j=0}^{\kappa}\sum_{l=0}^{\kappa-j}\frac{f_{a-\kappa+j-l}(X)R_{j,l}(X)}{(a-\kappa-l)!} = 0\,.$$

This completes the proof. ■

For the proof of Theorem 50, it will be useful to work only with generating functions and only translate back to distribution space at the end. We have the following equivalent result to the previous lemma. It can be proven in the exactly the same way.

Lemma 53 *Let $N(\mu_1, I+\Sigma_1), \dots, N(\mu_k, I+\Sigma_k)$ be Gaussians. Let $P_1(X), P_2(X), \dots, P_k(X)$ be polynomials in $X = (X_1, \dots, X_d)$ of degree at most m. Consider the generating function of the polynomial combination i.e.*

$$f(X,y) = P_1(Xy)e^{\mu_1(X)y+\frac{1}{2}\Sigma_1(X)y^2} + \cdots + P_k(Xy)e^{\mu_k(X)y+\frac{1}{2}\Sigma_k(X)y^2}$$

and let $f_0(X), f_1(X), \dots$ be the primary terms in the formal power series expansion of $f(y)$ i.e.

$$f(X,y) = \sum_{j=0}^{\infty}\frac{f_j(X)}{j!}y^j\,.$$

Let $\kappa = (m+1)(2^k-1)$. Then there are d-variate polynomials $R_{j,l}(X)$ for $0 \le j \le \kappa$ and $0 \le l \le \kappa - j$ such that

- *$R_{j,l}(X)$ is homogeneous of degree $\kappa - j + l$, $R_{\kappa,0}(X) = 1$*
- *$R_{j,l}(X)$ is $(O_{k,m}(1), O_{k,m}(1))$-simple with respect to $\{\mu_1(X), \Sigma_1(X), \dots, \mu_k(X), \Sigma_k(X)\}$*
- *For all integers $a \ge \kappa$,*

$$\sum_{j=0}^{\kappa}\sum_{l=0}^{\kappa-j}\frac{f_{a-\kappa+j-l}(X)R_{j,l}(X)}{(a-\kappa-l)!} = 0\,,$$

where undefined terms (i.e. negative factorials in the denominator) are treated as 0.

E.2. Observability When Components are All Very Close

Here, we deal with the special case when all pairs of components are close (within some small constant) in TV distance. In the next subsection, we will show how to reduce to this case.

We will first prove observability in an even simpler case where all of the components are within some small constant of isotropic.

Lemma 54 *Let $N(\mu_1, I+\Sigma_1), \dots, N(\mu_k, I+\Sigma_k)$ be Gaussians. Let $P_1(X), P_2(X), \dots, P_k(X)$ be polynomials of degree at most m. Let $\epsilon > 0$ be some parameter. Let $f(X,y)$ be the generating function of the polynomial combination and let $f_0(X), f_1(X), \dots$ be the primary terms in the formal power series expansion of $f(X,y)$. Then there exist constants c, C depending only on k and*

m with the following property. If $\|\mu_j\| \leq c$, $\|\Sigma_j\|_2 \leq c$ for all $j \in [k]$ and $\|v(f_j(X))\| \leq \epsilon$ for all $0 \leq j \leq C$, then

$$\|\chi(f(X,i))\|_1 \leq O_{m,k}(\epsilon)$$

where χ denotes the inverse adjusted characteristic function and $i = \sqrt{-1}$.

Proof By Lemma 53, we can write

$$\sum_{j=0}^{\kappa}\sum_{l=0}^{\kappa-j} \frac{f_{a-\kappa+j-l}(X)R_{j,l}(X)}{(a-\kappa-l)!} = 0\,.$$

for all $a \geq 2\kappa$ where the polynomials $R_{j,l}$ satisfy the properties in the statement of the lemma. This rearranges into

$$f_a(X) = -(a-\kappa)!\sum_{j=0}^{\kappa-1}\sum_{l=0}^{\kappa-j} \frac{f_{a-\kappa+j-l}(X)R_{j,l}(X)}{(a-\kappa-l)!}\,.$$

Using the triangle inequality and Claim 5,

$$\|v(f_a(X)\| \leq \sum_{j=0}^{\kappa-1}\sum_{l=0}^{\kappa-j} \frac{(a-\kappa)!}{(a-\kappa-l)!}\|v(f_{a-\kappa+j-l}(X))\|\,\|v(R_{j,l}(X)\|\,.$$

Let the sequence G_a be defined by $G_a = \|v(f_a(X)\|/\sqrt{a!}$. The above inequality implies

$$G_a \leq \sum_{j=0}^{\kappa-1}\sum_{l=0}^{\kappa-j} \frac{(a-\kappa)!\sqrt{(a-\kappa+j-l)!}}{(a-\kappa-l)!\sqrt{a!}}\|G_{a-\kappa+j-l}\|\,\|v(R_{j,l}(X)\|\,.$$

Note that

$$\frac{(a-\kappa)!\sqrt{(a-\kappa+j-l)!}}{(a-\kappa-l)!\sqrt{a!}} \leq \frac{(a-\kappa)!\sqrt{(a-2l)!}}{(a-\kappa-l)!\sqrt{a!}} = O_{m,k}(1)\,.$$

Also, note that by choosing c (the upper bound on $\|\mu_1\|, \dots, \|\mu_k\|, \|\Sigma_1\|_2, \dots, \|\Sigma_k\|_2$) sufficiently small, we can ensure that $\|v(R_{j,l}(X)\|$ is sufficiently small as a function of k, m for all $j < \kappa$. This is because the polynomials $R_{j,l}$ are homogeneous with positive degree and $(O_{m,k}(1), O_{m,k}(1))$-simple with respect to $\mu_1(X), \Sigma_1(X), \dots, \mu_k(X), \Sigma_k(X)$. Thus, as long as c is sufficiently small, we can ensure

$$G_a \leq 0.1\max(G_{a-1}, \dots, G_0)$$

for all $a \geq 2\kappa$. If we choose the constant $C > 2\kappa$, we may assume $G_0, G_1, \dots G_\kappa$ are all $O_{m,k}(\epsilon)$. Now we may apply Claim 17 to get

$$\|\chi(f(X,i))\|_1 \leq \sum_{j=0}^{\infty} \frac{\sqrt{j!}\,\|v(f_j)\|_2}{j!} = \sum_{j=0}^{\infty} G_j = O_{m,k}(\epsilon)$$

which completes the proof.

■

Now, by taking a suitable linear transformation, we can generalize the above result to when the components are in regular form and all pairs of components are sufficiently close in TV. There is some additional work to do because it is not immediately clear how taking a linear transformation affects things in generating function space. We prove the following result.

Corollary 55 *Let $N(\mu_1, I+\Sigma_1), \dots, N(\mu_k, I+\Sigma_k)$ be a set of Gaussians in (α, β)-regular form. Let $P_1(X), P_2(X), \dots, P_k(X)$ be polynomials of degree at most m. Let $\epsilon > 0$ be some parameter. Let $f(X, y)$ be the generating function of the polynomial combination and let $f_0(X), f_1(X), \dots$ be the primary terms in the formal power series expansion of $f(X, y)$.*

Then there exists a (sufficiently large) constant K depending only on k and m with the following property. If we have

$$\|\mu_j - \mu_{j'}\| \leq \frac{\beta^{-1}}{K}$$
$$\|\Sigma_j - \Sigma_j'\| \leq \frac{\beta^{-1}}{K}$$

for all $j, j' \in [k]$ and $\|v(f_j(X))\| \leq \epsilon$ for all $0 \leq j \leq K$, then

$$\|\chi(f(X, i))\|_1 \leq (2 + \alpha + \beta)^K \epsilon\,.$$

Proof We define $\mathcal{F} = \chi\left(f(X, i)\right)$. Let $\mu = \mu_1, \Sigma = \Sigma_1$. Let $L : \mathbb{R}^d \to \mathbb{R}^d$ denote the linear transformation $L(X) = (I + \Sigma)^{1/2} X + \mu$. Now write

$$\mathcal{F}(t) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} f(X, i) e^{-\frac{1}{2}\|X\|^2 - it \cdot X} dX\,.$$

Let $t = L(t')$ for some $t' \in \mathbb{R}^d$. Then

$$\begin{aligned}\mathcal{F}(t) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} f(X, i) e^{-\frac{1}{2}\|X\|^2 - it' \cdot (I+\Sigma)^{1/2} X - i\mu \cdot X} dX \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(\sum_{j=1}^k P_j(iX) e^{i\mu_j(X) - \frac{1}{2}\Sigma_j(X)} \right) e^{-\frac{1}{2}\|X\|^2 - it' \cdot (I+\Sigma)^{1/2} X - i\mu \cdot X} dX \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(\sum_{j=1}^k P_j(iX) e^{i(\mu_j - \mu) \cdot X - \frac{1}{2} X^T (\Sigma_j - \Sigma) X} \right) e^{-\frac{1}{2} X^T (I+\Sigma) X - it' \cdot (I+\Sigma)^{1/2} X} dX\end{aligned}$$

Let

$$g(X, y) = \left(\sum_{j=1}^k P_j(Xy) e^{(\mu_j - \mu) \cdot Xy + \frac{1}{2} X^T (\Sigma_j - \Sigma) X y^2} \right)\,.$$

Then we can substitute $X = (I + \Sigma)^{-1/2} Z$ and rewrite

$$\begin{aligned}\mathcal{F}(t) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} g(X, i) e^{-\frac{1}{2} X^T (I+\Sigma) X - it' \cdot (I+\Sigma)^{1/2} X} dX \\ &= \frac{\det(I + \Sigma)^{-1/2}}{(2\pi)^d} \int_{\mathbb{R}^d} g((I + \Sigma)^{-1/2} Z, i) e^{-\frac{1}{2}\|Z\|^2 - it' \cdot Z} dZ \\ &= \det(I + \Sigma)^{-1/2} \chi\left(g((I + \Sigma)^{-1/2} Z, i)\right)(t')\,.\end{aligned}$$

Recall that $t = L(t') = (I + \Sigma)^{1/2} t' + \mu$. Thus, we conclude

$$\|\mathcal{F}\|_1 = \left\|\chi\left(g((I + \Sigma)^{-1/2} Z, i)\right)\right\|_1\,. \tag{4}$$

We will now bound the RHS by applying Lemma 54.

Let C be the parameter from Lemma 54 (set in terms of k, m). Choosing K sufficiently large, we may assume that

$$\|v(f_j(X))\| \leq \epsilon \tag{5}$$

for all $j \leq C$. Note that

$$g(X, y) = f(X, y)e^{-\mu(X)y - \frac{1}{2}\Sigma(X)y^2}\,.$$

Let the primary terms of g be $g_0(X), g_1(X), \dots,$. We can also write the power series expansion of

$$e^{-\mu(X)y - \frac{1}{2}\Sigma(X)y^2} = \sum_{j=0}^{\infty} \frac{\left(-\mu(X)y - \frac{1}{2}\Sigma(X)y^2\right)^j}{j!} = \sum_{j=0}^{\infty} \frac{u_j(X)y^j}{j!}$$

for some polynomials u_j. Note that by Claim 5 and the fact that $\|\mu\|, \|\Sigma\|_2 \leq \alpha$, we have

$$\|v(u_j(X))\| \leq (2+\alpha)^{O_{m,k}(1)} \tag{6}$$

for $j \leq C$ (since $C = O_{m,k}(1)$). Now we have

$$\sum_{j=0}^{\infty} \frac{g_j(X)y^j}{j!} = \left(\sum_{j=0}^{\infty} \frac{f_j(X)y^j}{j!}\right)\left(\sum_{j=0}^{\infty} \frac{u_j(X)y^j}{j!}\right)\,.$$

Thus, after expanding and truncating to the first $C+1$ terms, we can use Claim 5 and equations (5) and (6) to deduce that for all j with $0 \leq j \leq C$,

$$\|v(g_j(X))\| \leq (2+\alpha)^{O_{m,k}(1)}\epsilon\,.$$

Now note that the primary terms of the function $g((I+\Sigma)^{-1/2}X, y)$ are exactly $g_j((I+\Sigma)^{-1/2}X)$. By Claim 7 we get for all j with $0 \leq j \leq C$,

$$\left\|v\left(g_j\left((I+\Sigma)^{-1/2}X\right)\right)\right\| \leq (2+\alpha+\beta)^{O_{m,k}(1)}\epsilon\,.$$

Now we need to check the remaining conditions of Lemma 54. Let

$$G'_j = N(\mu'_j, I + \Sigma'_j) = N\left((I+\Sigma)^{-1/2}(\mu_j - \mu), I + (I+\Sigma)^{-1/2}(\Sigma_j - \Sigma)(I+\Sigma)^{-1/2}\right)$$

for all $j \in [k]$. Note that we can write

$$g((I+\Sigma)^{-1/2}X, y) = \sum_{j \in [k]} P'_j(Xy)e^{\mu'_j(X)y + \frac{1}{2}\Sigma'_j(X)y^2}$$

for polynomials $P'_1, \dots, P'_k$ of degree at most m. By choosing K sufficiently large, we can ensure that the Gaussians $G'_1, \dots, G'_k$ satisfy the conditions of Lemma 54 (because $\|\mu_j - \mu\|, \|\Sigma_j - \Sigma\|_2$ will be sufficiently small). Thus, applying Lemma 54 and using (4) completes the proof. ■

E.3. Reducing to When Components are All Very Close

We will now deal with the case when the components are not necessarily all very close to each other. We will still assume that the components are in (α, β)-regular form for $\alpha, \beta \leq \mathrm{poly}(\log 1/\epsilon)$.

The way we will reduce to the case where the components are all very close is as follows. We show that we can partition the components $G_1, \dots, G_k$ into submixtures say $S_1, \dots, S_a \subset [k]$ such that

- Components in the same submixture are sufficiently close to apply Corollary 55
- Components in different submixtures are not too close

We then use differential operators to isolate each of these submixtures (relying on the second condition above) and deduce that if the Hermite moment polynomials of the entire mixture are close to 0, then the Hermite moment polynomials of each submixture are close to 0. We can then apply Corollary 55 on each submixture to complete the proof.

We will need a few additional definitions.

Definition 56 *Given Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$, we say a partition $S_1, \dots, S_a$ of $[k]$ is δ-separated if for any pair of components in different parts of the partition, say $i_1 \in S_{j_1}, i_2 \in S_{j_2}, j_1 \neq j_2$,*

$$\|\mu_{i_1} - \mu_{i_2}\|_2 + \|\Sigma_{i_1} - \Sigma_{i_2}\|_2 \geq \delta\,.$$

Definition 57 *Given Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$, we say a partition $S_1, \dots, S_a$ of $[k]$ is (δ_1, δ_2)-good for some parameters δ_1, δ_2 if for any pair of components in different parts of the partition, say $i_1 \in S_{j_1}, i_2 \in S_{j_2}, j_1 \neq j_2$,*

$$\|\mu_{i_1} - \mu_{i_2}\|_2 + \|\Sigma_{i_1} - \Sigma_{i_2}\|_2 \geq \delta_1\,.$$

and for any pair of components in the same part of the partition, say $i_1, i_2 \in S_{j_1}$,

$$\|\mu_{i_1} - \mu_{i_2}\|_2 + \|\Sigma_{i_1} - \Sigma_{i_2}\|_2 \leq \delta_2\,.$$

Claim 22 *Given Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$ and any parameter $\theta > 0$, there exists a $(\theta, k\theta)$-good partition of $[k]$.*

Proof Consider a graph on nodes $1, 2, \dots, k$ where two nodes i, j are connected if and only if

$$\|\mu_i - \mu_j\|_2 + \|\Sigma_i - \Sigma_j\|_2 \leq \theta\,.$$

Now let $S_1, \dots, S_a$ be the connected components in this graph. We claim this forms a $(\theta, k\theta)$-good partition. It is clear that any pair i, j in different parts of the partition satisfies

$$\|\mu_i - \mu_j\|_2 + \|\Sigma_i - \Sigma_j\|_2 \geq \theta\,.$$

On the other hand, for any pair in the same part of the partition, we can find a path between them in the graph, say $i, l_1, \dots, l_c, j$ and use the triangle inequality summed along the path to deduce

$$\|\mu_i - \mu_j\|_2 + \|\Sigma_i - \Sigma_j\|_2 \leq k\theta$$

which completes the proof. ■

Now we begin the main technical part. We first explain the intuition for using differential operators to isolate parts of the mixture. Let $\mathcal{D} = \partial - \mu(X) - \Sigma(X)y$ for some μ, Σ. Consider a Gaussian $N(\mu', I + \Sigma')$ and consider

$$\mathcal{D}\left(e^{\mu'(X)y + \frac{1}{2}\Sigma'(X)y^2}\right) .$$

Note that if μ, Σ are close to μ', Σ', then the result will be essentially 0. On the other hand, we can verify that if μ, Σ are not close to μ', Σ', then the result will be bounded away from 0. Thus, we can apply differential operators to essentially remove all components that are far away from $N(\mu', I + \Sigma')$, leaving only a submixture of components that are sufficiently close to each other. Over the next two claims, we will formalize this intuition by showing that given a submixture of close components, if we repeatedly apply far-away differential operators, we cannot accidentally zero-out the submixture.

Claim 23 *Consider a set of Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$. Let $P_1(X), \dots, P_k(X)$ be polynomials of degree at most m. Let $f(X, y)$ be the generating function of the polynomial combination and let $f_0, f_1, \dots$ be the primary terms in the formal power series expansion of f.*

Consider two parameters μ, Σ and assume that for all $j \in [k]$,

$$\delta \le \|\mu_j - \mu\|_2 + \|\Sigma_j - \Sigma\|_2 \le \delta^{-1} .$$

Let $g(X, y) = (\partial - (\mu(X) + \Sigma(X)y))f(X, y)$ where the partial derivative is taken with respect to y and let $g_0, g_1, \dots$ be the primary terms of g. Let K be a constant that is sufficiently large in terms of k, m. For any parameter $\epsilon > 0$, if for all $j \le K$

$$\|v(g_j(X))\|_2 \le \epsilon$$

then for all $j \le K$,

$$\|v(f_j(X))\|_2 \le (2 + \|\mu\| + \|\Sigma\|_2 + \delta^{-1})^{O_{k,m,K}(1)}\epsilon .$$

Proof Consider

$$h(X, y) = g(X, y)e^{-\mu(X)y - \frac{1}{2}\Sigma(X)y^2} .$$

Note that

$$h(X, y) = \partial_y\left(f(X, y)e^{-\mu(X)y - \frac{1}{2}\Sigma(X)y^2}\right) .$$

Let $F(X, y) = f(X, y)e^{-\mu(X)y - \frac{1}{2}\Sigma(X)y^2}$ and let $F_0, F_1, \dots$ be its primary terms. The primary terms of h are $F_1, F_2, \dots$ (i.e. the same but shifted down by one).

First note that since $h(X, y) = g(X, y)e^{-\mu(X)y - \frac{1}{2}\Sigma(X)y^2}$, we have for all $1 \le j \le K + 1$

$$\|v(F_j(X))\|_2 \le (2 + \|\mu\| + \|\Sigma\|_2)^{O_K(1)}\epsilon . \tag{7}$$

To see this, it suffices to consider the product of the power series for $g(X,y)$ and $e^{-\mu(X)y-\frac{1}{2}\Sigma(X)y^2}$ after truncating both series at y^K (dropping all terms with higher powers of y). By Claim 5, the first $K+1$ primary terms in the power series expansion of $e^{-\mu(X)y-\frac{1}{2}\Sigma(X)y^2}$ have coefficient norm at most $(2+\|\mu\|+\|\Sigma\|_2)^{O_K(1)}$. Thus, when we expand out the product of $g(X,y)$ and $e^{-\mu(X)y-\frac{1}{2}\Sigma(X)y^2}$, we use Claim 5 again to deduce that the resulting primary terms will all have coefficient norm at most $(2+\|\mu\|+\|\Sigma\|_2)^{O_K(1)}\epsilon$.

Now we will argue about $F_0(X)$ (which is just a constant). Define the following differential operators for $j \in [k]$,

$$\Delta_j = \partial - ((\mu_j(X) - \mu(X)) + (\Sigma_j(X) - \Sigma(X))y)\,.$$

Now consider the differential operator

$$\Delta = \Delta_k^{(m+1)2^{k-1}} \Delta_{k-1}^{(m+1)2^{k-2}} \dots \Delta_1^{m+1}\,.$$

We know by Claim 21 that $\Delta(F(X,y)) = \Delta(f(X,y)e^{-\mu(X)y-\frac{1}{2}\Sigma(X)y^2}) = 0$. On the other hand, we may use Claim 18 to expand Δ in the form

$$\Delta = \partial^{(m+1)(2^k-1)} + R_{(m+1)(2^k-1)-1}(X,y)\partial^{(m+1)(2^k-1)-1} + \dots + R_1(X,y)\partial + R_0(X,y)\,.$$

for some polynomials $R_0, R_1, \dots$. Let $\kappa = (m+1)(2^k-1)$. We have

$$R_0(X,y)F(X,y) = -\left(\sum_{j=1}^{\kappa} R_j(X,y)\partial^{\kappa-j}(F(X,y))\right)$$

which is equivalent to

$$R_0(X,y)\sum_{l=0}^{\infty}\frac{F_l(X)y^l}{l!} = -\left(\sum_{j=1}^{\kappa} R_j(X,y)\left(\sum_{l=0}^{\infty}\frac{F_{l+j}(X)y^l}{l!}\right)\right)\,.$$

The key observation is that $F_0(X)$ appears on the LHS but not the RHS i.e. we may write

$$R_0(X,y)F_0(X) = -\left(\sum_{j=1}^{\kappa} R_j(X,y)\left(\sum_{l=0}^{\infty}\frac{F_{l+j}(X)y^l}{l!}\right)\right) - R_0(X,y)\sum_{l=1}^{\infty}\frac{F_l(X)y^l}{l!}\,. \quad (8)$$

Now by Claim 20, we have

$$\|v_y(R_0(X,y))\| \geq (0.1\delta)^{O_{k,m}(1)}\,.$$

Note that Claim 18 and Claim 5 give an upper bound on the coefficient norm of $R_j(X,y)$ for all $0 \leq j \leq \kappa$. In particular, we get

$$\|v_y(R_j(X,y))\| \leq (2\delta^{-1})^{O_{k,m}(1)}\,.$$

Now, by combining with (7) we can upper bound the coefficient norm of the first κ terms in the power series of the RHS of (8). Since $R_0(X,y)$ has degree at most κ, as long as $K > 10\kappa$, we get

$$\|v(F_0(X))\| \leq (2+\|\mu\|+\|\Sigma\|_2+\delta^{-1})^{O_{K,k,m}(1)}\epsilon\,.$$

Now we have an upper bound on the coefficient norm of all of $F_0(X), \dots, F_K(X)$. Note that

$$f(X, y) = F(X, y)e^{\mu(X)y + \frac{1}{2}\Sigma(X)y^2}$$

and we can expand out the product of the power series of $F(X, y)$ and $e^{\mu(X)y+\frac{1}{2}\Sigma(X)y^2}$, using the same argument as before, to get that for all j with $0 \leq j \leq K$,

$$\|v(f_j(X))\| \leq (2 + \|\mu\| + \|\Sigma\|_2 + \delta^{-1})^{O_{K,k,m}(1)}\epsilon\,.$$

■

By repeatedly applying the previous claim, we get the following.

Claim 24 *Consider a set of Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$ in (α, β)-regular form and let $P_1(X), \dots, P_k(X)$ be polynomials of degree at most m. Let $f(X, y)$ be the generating function of the polynomial combination and let $f_0, f_1, \dots$ be the primary terms in the formal power series expansion of f.*

Let $S_1, \dots, S_a$ be a δ-separated partition of $[k]$ for some constant $\delta < 1$. For each $l \in [a]$, let $f^{(l)}(X, y)$ be the generating function of the polynomial combination of only the Gaussians in S_l and let $f_0^{(l)}, f_1^{(l)}, \dots$ be the primary terms in the formal power series expansion of $f^{(l)}$.

For any constant K that is sufficiently large in terms of k, m, There exists a constant $C_{k,K,m}$ depending on k, K, m such that the following holds. If for all $j \leq K$ we have

$$\|v(f_j(X))\| \leq \epsilon$$

then for all $l \in [a], j \leq K - (m + 1)2^k$,

$$\left\|v(f_j^{(l)}(X))\right\| \leq \epsilon \left(2 + \alpha + \delta^{-1}\right)^{C_{k,K,m}}\,.$$

Proof Without loss of generality $S_1 = \{1, 2, \dots, t\}$ for some $t \leq k$. Recall that for each $j \in [k]$, we use $\mathcal{D}_j$ to denote the differential operator $\partial - (\mu_j(X) + \Sigma_j(X)y)$. Now consider the differential operator

$$\mathcal{D}^{(1)} = \mathcal{D}_k^{2^{k-t-1}(m+1)} \cdots \mathcal{D}_{t+2}^{2(m+1)}\mathcal{D}_{t+1}^{m+1}\,.$$

Note that by Claim 18 and Claim 5, we can rewrite

$$\mathcal{D}^{(1)} = \partial^\kappa + R_{\kappa-1}(X, y)\partial^{\kappa-1} + \cdots + R_0(X, y)$$

where $\kappa = (m + 1)(2^{k-t} - 1)$ and for all j,

$$\|v_y(R_j(X, y))\| \leq (2 + \alpha)^{O_{k,m}(1)}\,.$$

Now consider the power series expansion of $\mathcal{D}^{(1)}f$ and let $d_0(X), d_1(X), \dots$ be its primary terms. Since the differential operator $\mathcal{D}^{(1)}$ has degree at most $(m + 1)2^k$, the assumption in the statement of the claim implies that for all $j \leq K - (m + 1)2^k$,

$$\|v(d_j(X))\| \leq (2 + \alpha)^{O_{k,m,K}(1)}\,\epsilon\,.$$

On the other hand, note that

$$\mathcal{D}^{(1)} f = \mathcal{D}^{(1)} f^{(1)}$$

since by Claim 21, the operator $\mathcal{D}^{(1)}$ zeros out all of the other components. We may now repeatedly apply Claim 23 (since $\mathcal{D}^{(1)}$ factors as a composition of linear differential operators) and use the fact that the partition $S_1, \dots, S_a$ is δ-separated to deduce that as long as K is sufficiently large in terms of k, m, we have for all $j \leq K - (m+1)2^k$,

$$\left\| v(f_j^{(1)}(X)) \right\| \leq \left(2 + \alpha + \delta^{-1}\right)^{O_{k,m,K}(1)} \epsilon \,.$$

Since the initial choice of $l = 1$ was arbitrary, this completes the proof. ■

We can now prove the main theorem of this section. The statement below is stated in terms of generating functions. We will translate it into distribution space and prove Theorem 50 immediately afterwards.

Theorem 58 *Consider a set of Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$ in (α, β)-regular form and let $P_1(X), \dots, P_k(X)$ be polynomials of degree at most m. Let $f(X, y)$ be the generating function of the polynomial combination and let $f_0, f_1, \dots$ be the primary terms in the formal power series expansion of f.*

There exists a constant C depending only on k, m such that the following holds. If for all j with $0 \leq j \leq C$, we have

$$\|v(f_j(X))\| \leq \epsilon$$

then we must have

$$\|\chi(f(X, i))\|_1 \leq (2 + \alpha + \beta)^C \epsilon \,.$$

Proof Let K be the constant in Corollary 55. Recall that K is set as a function of k, m. Let $\theta = \frac{\beta^{-1}}{2kK}$. By Claim 22, there is a $(\theta, k\theta)$-good partition of $[k]$, say $S_1, \dots, S_a$. For each $l \in [a]$, let

$$f^{(l)}(y) = \sum_{j \in S_l} P_j(Xy) e^{\mu_j(X)y + \frac{1}{2}\Sigma_j(X)y^2} \,.$$

and let $f_0^{(l)}, f_1^{(l)}, \dots$ be the primary terms in the expansion of $f^{(l)}$ as a power series in y.

By Claim 24, as long as C is sufficiently large in terms of k, m, we have for all $0 \leq j \leq K$

$$\left\| v(f_j^{(l)}(X)) \right\| \leq \epsilon (2 + \alpha + \beta)^{O_{m,k}(1)} \,.$$

Now by Corollary 55, since the partition is such that all pairs of components in the same part are close, we get that

$$\left\| \chi\left(f^{(l)}(X, i)\right) \right\|_1 \leq \epsilon (2 + \alpha + \beta)^{O_{m,k}(1)} \,.$$

Finally, since

$$f = f^{(1)} + \dots + f^{(a)}$$

we get

$$\|\chi\left(f(X, i)\right)\|_1 \leq (2 + \alpha + \beta)^{O_{m,k}(1)} \epsilon$$

which completes the proof. ■

To prove Theorem 50, it suffices to translate the above theorem from generating functions back to distributions.

Proof [Proof of Theorem 50] By Corollary 41, there are polynomials $P_1, \dots, P_k$ of degree at most m such that

$$f(X,y) = \sum_{j=0}^{\infty} \frac{1}{j!} h_{j,g}(X) y^j = P_1(Xy) e^{\mu_1(X)y + \frac{1}{2}\Sigma_1(X)y^2} + \dots + P_k(Xy) e^{\mu_k(X)y + \frac{1}{2}\Sigma_k(X)y^2} \,.$$

Now, applying Theorem 58 to the expression on the RHS, we have

$$\|\chi(f(X,i))\|_1 \leq (2 + \alpha + \beta)^{O_{m,k}(1)} \epsilon \,.$$

Finally, Fact 1 and Claim 14 imply that $g = \chi(f(X,i))$ so we are done. ■

Appendix F. Hermite Moment Polynomials of a Single Gaussian: Tail Bounds and other Properties

Note that the Hermite moment polynomials of a distribution are given by $\mathbb{E}_z[H_m(X,z)]$ for z drawn from that distribution. The way we estimate these Hermite moment polynomials from samples will be by robustly estimating the mean of the distribution $H_m(X,z)$. In this section, our goal is to understand properties of the distribution of $H_m(X,z)$ for z drawn from a single Gaussian of the form $N(\mu, I + \Sigma)$. Later, in Section G, we will use this to deduce that we can estimate $H_m(X,z)$ to within nearly optimal error even when z is drawn from a regular-form mixture of Gaussians.

We will need a few definitions.

Definition 59 *Given two sets of d variables $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)})$ and $X^{(2)} = (X_1^{(2)}, \dots, X_d^{(2)})$ and a polynomial $P(X^{(1)}, X^{(2)})$, for integers m_1, m_2, the degree-(m_1, m_2)-part of P consists of the monomials of P that have total degree m_1 in $X^{(1)}$ and total degree m_2 in $X^{(2)}$.*

Definition 60 *Consider two sets of formal variables, say*

$$X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)}), X^{(2)} = (X_1^{(2)}, \dots, X_d^{(2)}) \,.$$

Let $G = N(\mu, I + \Sigma)$ be a Gaussian. We say the fundamental polynomials (with respect to G) are

$$\left\{ \mu\left(X^{(1)}\right), \mu\left(X^{(2)}\right), \Sigma\left(X^{(1)}\right), \Sigma\left(X^{(2)}\right), \left(X^{(1)}\right)^T (I + \Sigma) X^{(2)} \right\}$$

F.1. Covariance of Multivariate Hermite Polynomials

First, we analyze the covariance of $H_m(X,z)$ for z drawn from a Gaussian $N(\mu, I + \Sigma)$.

Claim 25 *Let $G = N(\mu, I + \Sigma)$ be a Gaussian. Let m be a positive integer. Consider two sets of d variables $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)})$ and $X^{(2)} = (X_1^{(2)}, \dots, X_d^{(2)})$. Then the expression*

$$\mathbb{E}_{z \sim G}[H_m(X^{(1)}, z) H_m(X^{(2)}, z)]$$

(which is a formal polynomial in $2d$ variables) can be computed as follows:

1. *Consider the power series expansion of*

$$\exp\left(y\mu(X^{(1)}+X^{(2)})+\frac{1}{2}y^2\Sigma(X^{(1)}+X^{(2)})+y^2X^{(1)}\cdot X^{(2)}\right)=\sum_{j=0}^{\infty}\frac{Q_j(X^{(1)},X^{(2)})y^j}{j!}$$

2. *Take* $\binom{2m}{m}^{-1}$ *times the degree* (m,m) *part of* $Q_{2m}(X^{(1)},X^{(2)})$

Proof We will evaluate $\mathbb{E}_{z\sim G}[H_m(X^{(1)},z)H_m(X^{(2)},z)]$ using generating functions. Let

$$F=\left(\sum_{m=0}^{\infty}\frac{1}{m!}H_m(X^{(1)},z)y^m\right)\left(\sum_{m=0}^{\infty}\frac{1}{m!}H_m(X^{(2)},z)y^m\right).$$

Note that $H_m(X,z)$ is homogeneous and of degree m in X. Thus, to compute $H_m(X^{(1)},z)H_m(X^{(2)},z)$, it suffices to extract the degree-(m,m) part of the coefficient of y^{2m} in the power series expansion of F.

Now by Claim 11 we may write

$$\begin{aligned}
&\mathbb{E}_{z\sim G}[F]=\mathbb{E}_{z\sim G}\left[\exp\left(yz\cdot(X^{(1)}+X^{(2)})-\frac{1}{2}y^2(X^{(1)}\cdot X^{(1)}+X^{(2)}\cdot X^{(2)})\right)\right]\\
&=C\int\exp\left(-\frac{1}{2}(z-\mu)^T(I+\Sigma)^{-1}(z-\mu)+yz\cdot(X^{(1)}+X^{(2)})-\frac{1}{2}y^2(X^{(1)}\cdot X^{(1)}+X^{(2)}\cdot X^{(2)})\right)\\
&=C\int\exp\Bigg(-\frac{1}{2}\left(z-\mu-y(I+\Sigma)(X^{(1)}+X^{(2)})\right)^T(I+\Sigma)^{-1}\left(z-\mu-y(I+\Sigma)(X^{(1)}+X^{(2)})\right)\\
&\quad+y\mu\cdot(X^{(1)}+X^{(2)})+\frac{1}{2}y^2\left(\left(X^{(1)}\right)^T\Sigma X^{(1)}+\left(X^{(2)}\right)^T\Sigma X^{(2)}+2\left(X^{(1)}\right)^T(I+\Sigma)X^{(2)}\right)\Bigg)\\
&=\exp\left(y\mu(X^{(1)}+X^{(2)})+\frac{1}{2}y^2\Sigma(X^{(1)}+X^{(2)})+y^2X^{(1)}\cdot X^{(2)}\right).
\end{aligned}$$

In the above, C denotes the normalization constant for a Gaussian with covariance $I+\Sigma$. Next, expanding the above as a power series, we know that the degree (m,m) part of $Q_{2m}(X^{(1)},X^{(2)})$ is equal to

$$\mathbb{E}\left[\frac{(2m)!}{m!\cdot m!}H_m(X^{(1)},z)H_m(X^{(2)},z)\right]$$

from which we immediately get the desired conclusion. ■

As a corollary to the above, we get the following upper bound on the covariance of $v(H_m(X,z))$ for a single Gaussian. To do this, we rely on the symmetric tensorization (recall definition 33) and its properties to relate the expression $\mathbb{E}_{z\sim G}[H_m(X^{(1)},z)H_m(X^{(2)},z)]$ to the covariance of the vector $v(H_m(X,z))$.

Claim 26 *Let* $G=N(\mu,I+\Sigma)$ *be a Gaussian. Let* m *be a positive integer. Let* Σ_{H_m} *be the covariance of* $v(H_m(X,z))$ *for* z *drawn from* G. *Then*

$$\Sigma_{H_m}\leq E_{z\sim G}\left[v(H_m(X,z))\otimes v(H_m(X,z))\right]\leq(m(1+\|\mu\|_2+\|\Sigma\|_2))^{O(m)}I$$

where I *on the RHS denotes the identity matrix of the appropriate size.*

Proof Consider two sets of d variables $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)})$ and $X^{(2)} = (X_1^{(2)}, \dots, X_d^{(2)})$. Let

$$P(X^{(1)}, X^{(2)}) = \mathbb{E}_{z\sim G}[H_m(X^{(1)}, z) H_m(X^{(2)}, z)] \,.$$

Let $T = T_{\mathsf{sym}}(P)$. Note that by Claim 9,

$$T = \mathbb{E}_{z\sim G}[v(H_m(X^{(1)}, z)) \otimes v(H_m(X^{(2)}, z))] \,.$$

On the other hand, by Claim 25, $P(X^{(1)}, X^{(2)})$ can be computed by considering the power series expansion

$$\exp\left(y\mu(X^{(1)} + X^{(2)}) + \frac{1}{2}y^2\Sigma(X^{(1)} + X^{(2)}) + y^2 X^{(1)} \cdot X^{(2)}\right) = \sum_{j=0}^{\infty} \frac{Q_j(X^{(1)}, X^{(2)}) y^j}{j!} \,.$$

and taking $\binom{2m}{m}^{-1}$ times the degree (m, m) part of Q_{2m}. We will use $Q_{m,m}$ to denote the degree (m, m) part of Q_{2m}. Write

$$\begin{aligned} Q &= y\mu(X^{(1)} + X^{(2)}) + \frac{1}{2}y^2\Sigma(X^{(1)} + X^{(2)}) + y^2 X^{(1)} \cdot X^{(2)} \\ &= y\mu^T(X^{(1)} + X^{(2)}) + y^2(X^{(1)})^T(I + \Sigma)X^{(2)} + \frac{1}{2}y^2\left((X^{(1)})^T\Sigma X^{(1)} + (X^{(2)})^T\Sigma X^{(2)}\right) \,. \end{aligned}$$

Now we may write

$$\exp(Q) = 1 + Q + \frac{Q^2}{2!} + \dots \,.$$

Let $\mathcal{S}$ be the set of fundamental polynomials of $G = N(\mu, I + \Sigma)$. Note that Q is a sum of $O(1)$ polynomials from among $\mathcal{S}$. Furthermore, each of these polynomials is homogeneous in each of the sets of variables $X^{(1)}, X^{(2)}$. Thus, we can expand each of the terms $Q, Q^2, \dots$ as a sum of products of elements of $\mathcal{S}$. Now by Claim 25, we can obtain $P(X^{(1)}, X^{(2)})$ by discarding all of the products that do not have the proper degrees (degree exactly m in each of the subsets of variables $X^{(1)}, X^{(2)}$). Since Q is a sum of $O(1)$ polynomials from $\mathcal{S}$, each with degree 1 or 2, and we only keep terms with total degree $2m$, we deduce that the number of terms (and all of the coefficients in front of the terms) that we keep are $m^{O(m)}$. Thus, $P(X^{(1)}, X^{(2)})$ is $(m^{O(m)}, 2m)$-simple with respect to $\mathcal{S}$.

It now suffices to bound the operator norms of the symmetric tensorizations of each of the individual products of fundamental polynomials. Consider a product of fundamental polynomials $P_1P_2 \cdots P_a$ for some $a \le 2m$. By Claim 10,

$$\begin{aligned} \left\|T_{\mathsf{sym}}(P_1P_2\cdots P_a)\right\|_{\mathsf{op}} &\le \left\|T_{\mathsf{sym}}(P_1)\right\| \cdots \left\|T_{\mathsf{sym}}(P_a)\right\| \\ &\le \Big((\|\mu\|_2 + 1)(\|\Sigma\|_2 + 1)(\|I + \Sigma\|_{\mathsf{op}} + 1)\Big)^{O(m)} \le (1 + \|\mu\|_2 + \|\Sigma\|_2)^{O(m)} \,. \end{aligned}$$

Combining this inequality with the fact that $P(X^{(1)}, X^{(2)})$ is $\left(m^{O(m)}, 2m\right)$-simple with respect to $\mathcal{S}$ gives that

$$T \le (m(1 + \|\mu\|_2 + \|\Sigma\|_2))^{O(m)} I \,.$$

Also, clearly $\Sigma_{H_m} \le T$ so we are done. ■

We will also need a lower bound on the covariance of the Hermite polynomials. This lower bound will hold when $G = N(\mu, I + \Sigma)$ is within a sufficiently small constant of isotropic.

Claim 27 *Let $G = N(\mu, I + \Sigma)$ be a Gaussian. Let m be a positive integer. There exists a constant c_m depending only on m such that if $\|\mu\|, \|\Sigma\|_2 \leq c_m$, then Σ_{H_m}, the covariance of $v(H_m(X, z))$ for z drawn from G, satisfies*

$$\Sigma_{H_m} \geq \frac{1}{2} I \,.$$

Proof Consider two sets of d variables $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)})$ and $X^{(2)} = (X_1^{(2)}, \dots, X_d^{(2)})$. Let

$$P(X^{(1)}, X^{(2)}) = \mathbb{E}_{z \sim G}[H_m(X^{(1)}, z) H_m(X^{(2)}, z)] \,.$$

Let $T = T_{\mathsf{sym}}(P)$. Note that

$$T = \mathbb{E}_{z \sim G}[v(H_m(X^{(1)}, z)) \otimes v(H_m(X^{(2)}, z))] \,.$$

On the other hand, recall that by Claim 25, $P(X^{(1)}, X^{(2)})$ can be computed by considering the power series expansion

$$\exp\left(y\mu(X^{(1)} + X^{(2)}) + \frac{1}{2} y^2 \Sigma(X^{(1)} + X^{(2)}) + y^2 X^{(1)} \cdot X^{(2)}\right) = \sum_{j=0}^{\infty} \frac{Q_j(X^{(1)}, X^{(2)}) y^j}{j!} \,.$$

and taking $\binom{2m}{m}^{-1}$ times the degree (m, m) part of Q_{2m}. We will use $Q_{m,m}$ to denote the degree (m, m) part of Q_{2m}.

The key observation now is that if $\mu = 0, \Sigma = 0$, then the LHS of the above is just $\exp(y^2 X^{(1)} \cdot X^{(2)})$. In this case, $Q_{m,m}$ would be $(2m)!/m! \cdot (X^{(1)} \cdot X^{(2)})^m$.

We now show that by choosing c_m sufficiently small, we can ensure that $Q_{m,m}$ does not change by too much. Let us write the power series expansion

$$\exp\left(y\mu(X^{(1)} + X^{(2)}) + \frac{1}{2} y^2 \Sigma(X^{(1)} + X^{(2)})\right) = \sum_{j=0}^{\infty} \frac{R_j(X^{(1)}, X^{(2)}) y^j}{j!} \,.$$

For each j, let $R_{j,j}(X^{(1)}, X^{(2)})$ denote the degree (j, j) part of R_{2j}. Next note that

$$\frac{Q_{m,m}(X^{(1)}, X^{(2)})}{(2m)!} = \sum_{j=0}^{m} \frac{(X^{(1)} \cdot X^{(2)})^{m-j} R_{j,j}(X^{(1)}, X^{(2)})}{(m-j)!(2j)!} \,.$$

Consider the expression $(X^{(1)} \cdot X^{(2)})^{m-j} R_{j,j}(X^{(1)}, X^{(2)})$ for a fixed j. Note that by choosing c_m sufficiently small, we can ensure, using Claim 10, that

$$\left\| T_{\mathsf{sym}}(R_{j,j}(X^{(1)}, X^{(2)})) \right\|_2$$

is bounded by a sufficiently small function of m for $1 \leq j \leq m$. Next, note that

$$T_{\mathsf{sym}}((X^{(1)} \cdot X^{(2)})^{m-j}) = I$$

where I is the identity matrix of the appropriate size. Thus, by Claim 10. we can ensure that

$$\left\| T_{\mathsf{sym}}\left((X^{(1)} \cdot X^{(2)})^{m-j} R_{j,j}(X^{(1)}, X^{(2)})\right) \right\|_{\mathsf{op}}$$

is bounded by a sufficiently small function of m for all $1 \leq j \leq m$. Thus, if we let

$$\begin{aligned} A &= T_{\mathsf{sym}}\left(Q_{m,m}(X^{(1)}, X^{(2)}) - \frac{(2m)!}{m!}(X^{(1)} \cdot X^{(2)})^m\right) \\ &= T_{\mathsf{sym}}\left((2m)! \sum_{j=1}^{m} \frac{(X^{(1)} \cdot X^{(2)})^{m-j} R_{j,j}(X^{(1)}, X^{(2)})}{(m-j)!(2j)!}\right) \end{aligned}$$

then we can ensure that $\|A\|_{\mathsf{op}}$ is bounded by a sufficiently small function of m.

However, the symmetric tensorization of $(X^{(1)} \cdot X^{(2)})^m$ is exactly the identity matrix I. Thus, we have

$$T \geq 0.9I$$

where recall

$$T = \mathbb{E}_{z\sim G}[v(H_m(X^{(1)}, z)) \otimes v(H_m(X^{(2)}, z))] = T_{\mathsf{sym}}\left(\binom{2m}{m}^{-1} Q_{m,m}(X^{(1)}, X^{(2)})\right) .$$

To bound the covariance of $v(H_m(X, z))$, it remains to compute

$$\mathbb{E}_{z\sim G}[v(H_m(X, z))] \otimes \mathbb{E}_{z\sim G}[v(H_m(X, z))] .$$

However, by Claim 12, $\mathbb{E}_{z\sim G}[H_m(X, z)] = h_{m,G}(X)$ is exactly the coefficient of $y^m/m!$ in the power series expansion of

$$e^{\mu(X)y + \frac{1}{2}\Sigma(X)y^2} .$$

As long as c_m is sufficiently small, we can use Claim 5 to get that

$$\|\mathbb{E}_{z\sim G}[v(H_m(X, z))] \otimes \mathbb{E}_{z\sim G}[v(H_m(X, z))]\|_2 = \|v(h_{m,G}(X))\|^2 \leq 0.1 .$$

Putting everything together, we get that

$$\Sigma_{H_m} = T - \mathbb{E}_{z\sim G}[v(H_m(X, z))] \otimes \mathbb{E}_{z\sim G}[v(H_m(X, z))] \geq 0.5I .$$

■

F.2. Tail Bounds and Stability Bounds

Now we prove that the distribution of $H_m(X, z)$ for z drawn from a Gaussian $G = N(\mu, I + \Sigma)$ has exponential tail decay. This will let us obtain tight stability bounds for samples drawn from this distribution. The stability bounds will then be plugged into existing algorithms for robust mean estimation (see e.g. Diakonikolas and Kane (2019) for an explanation of stability and its use in robust mean estimation).

The proofs in this section are mostly standard and many of them are deferred to Appendix J. First we need a tail bound on the distribution of $H_m(X, z)$.

Lemma 61 *Let $G = N(\mu, I + \Sigma)$. Consider the vector $v(H_m(X, z))$ for $z \sim G$. Let u be any unit vector with the same dimensionality. There are positive constants c_m, C_m depending only on m such that for any real number $t > 1$*

$$\Pr\left[|v(H_m(X, z)) \cdot u| \geq t(2 + \|\mu\|_2 + \|\Sigma\|_2)^{C_m}\right] \leq e^{-t^{c_m}}$$

Proof We will use Claim 26 and Claim 4 to bound

$$\mathbb{E}_{z\sim G}\left[(v(H_m(X,z))\cdot u)^k\right]$$

for some appropriately chosen even integer k and then use Markov's inequality. First, Claim 26 implies

$$\mathbb{E}_{z\sim G}\left[(v(H_m(X,z))\cdot u)^2\right] \leq (m(1+\|\mu\|_2+\|\Sigma\|_2))^{O(m)}\,.$$

Now note that for a fixed u, $v(H_m(X,z))\cdot u$ is a polynomial in z of degree at most m. Thus, by Claim 4, we get that for even integers k,

$$\mathbb{E}_{z\sim G}\left[(v(H_m(X,z))\cdot u)^k\right] \leq (mk(1+\|\mu\|_2+\|\Sigma\|_2))^{O(mk)}\,.$$

Now by Markov's inequality,

$$\Pr\left[|v(H_m(X,z))\cdot u| \geq t(2+\|\mu\|_2+\|\Sigma\|_2)^{C_m}\right] \leq \frac{(mk(1+\|\mu\|_2+\|\Sigma\|_2))^{O(mk)}}{t^k(2+\|\mu\|_2+\|\Sigma\|_2)^{C_mk}}\,.$$

Choosing C_m to be sufficiently large in terms of m and $k=t^{c_m}$ for some sufficiently small positive constant c_m depending only on m gives the desired inequality. ■

Now that we have shown that the distribution of $H_m(X,z)$ exhibits exponential tail decay in all directions, we can prove finite sample concentration inequalities. First we prove a concentration inequality in 1D, stating that for a set of samples from a distribution with exponential tail decay, with high probability, the empirical mean of any $(1-\epsilon)$ fraction of the samples is within $\widetilde{O}(\epsilon)$ of the true mean.

Claim 28 *Let $\mathcal{D}$ be a distribution on $\mathbb{R}$ and $0<c<1$ be a positive constant such that for all real numbers $t>1$,*

$$\Pr_{x\sim\mathcal{D}}[|x|\geq t] \leq e^{-t^c}\,.$$

Let $\epsilon<1/2$ and d be parameters. Given a set S of $n\geq (d/\epsilon)^{10^5/c}$ independent samples from $\mathcal{D}$, with probability at least $1-e^{-(8d/\epsilon)^2}$, any subset $S'\subseteq S$ of size at least $(1-\epsilon)n$ satisfies

$$\left|\mu_{\mathcal{D}} - \frac{1}{|S'|}\sum_{x\in S'} x\right| \leq \epsilon\log^{1/c}(1/\epsilon)\left(\frac{10^2}{c}\right)^{10/c}\,.$$

Proof See Appendix J

■

Now, we are ready to introduce the definition of stability of a set of samples in $\mathbb{R}^d$. The definition below is standard in robust statistics literature (see e.g. Diakonikolas and Kane (2019)).

Definition 62 *For $\epsilon>0$ and $\delta\geq\epsilon$, we say a finite set $S\subset\mathbb{R}^d$ is (ϵ,δ)-stable with respect to a distribution $\mathcal{D}$ if for every unit vector $v\subset\mathbb{R}^d$ and every subset $S'\subseteq S$ of size at least $(1-\epsilon)|S|$ we have*

$$\left|v\cdot\left(\mu_{\mathcal{D}}-\frac{1}{S'}\sum_{x\in S'}x\right)\right| \leq \delta$$

$$\left|\frac{1}{S'}\sum_{x\in S'}(v\cdot(x-\mu_{\mathcal{D}}))^2-1\right| \leq \frac{\delta^2}{\epsilon}$$

Note that the previous definition makes sense for a distribution $\mathcal{D}$ that has covariance I. However, we will need to work with distributions with unknown covariance. We make the following analogous definition of stability for distributions with unknown covariance.

Definition 63 *For $\epsilon > 0$ and $\delta \geq \epsilon$, we say a finite set $S \subset \mathbb{R}^d$ is (ϵ, δ)-pseudo-stable with respect to a distribution $\mathcal{D}$ if for every unit vector $v \subset \mathbb{R}^d$ and every subset $S' \subseteq S$ of size at least $(1-\epsilon)|S|$ we have*

$$\left| v \cdot \left(\mu_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} x \right) \right| \leq \delta$$

$$\left| v^T \left(\Sigma_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} (x - \mu_{\mathcal{D}})(x - \mu_{\mathcal{D}})^T \right) v \right| \leq \frac{\delta^2}{\epsilon}$$

Remark 64 *Note that if the covariance matrix of $\mathcal{D}$, $\Sigma_{\mathcal{D}}$, is well-conditioned, then a set of samples that is (ϵ, δ)-pseudo-stable can be transformed into a set of samples that is $(O(\epsilon), O(\delta))$-stable by applying a suitable linear transformation. However, if the covariance matrix is not well-conditioned, then the definitions of stability and pseudo-stability are incomparable.*

We now prove the main result of this section. It states that for a set S of samples drawn from the distribution of $H_m(X, z)$ for $z \sim \mathcal{M}$ where $\mathcal{M}$ is a regular-form mixture of Gaussians, S is $(\epsilon, \widetilde{O}(\epsilon))$ pseudo-stable with high probability. The proof involves combining Lemma 61 and Claim 28 and union bounding over a sufficiently fine discrete net over the set of all possible directions. The details are deferred to Appendix J.

Claim 29 *Consider a set of Gaussians $G_1 = N(\mu_1, I + \Sigma_1), \dots, G_k = N(\mu_k, I + \Sigma_k)$ in $\mathbb{R}^d$ and assume they are in (α, β)-regular form.*

Consider the mixture $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ (where $w_1, \dots, w_k$ are nonnegative weights summing to 1). Let m be a positive integer and let $n > \mathrm{poly}_{k,m}(d/\epsilon)$ for some sufficiently large polynomial. Let $\mathcal{D}$ be the distribution of $v(H_m(X, z))$ for z drawn from $\mathcal{M}$. Consider a set S of n such samples (drawn independently) $x_1 = v(H_m(X, z_1)), \dots, x_n = v(H_m(X, z_n))$. This set of n samples is (ϵ, δ)-pseudo-stable with

$$\delta = \epsilon \left(2 + \alpha + \beta + \log(1/\epsilon)\right)^{O_{m,k}(1)}$$

with probability $1 - e^{-10d/\epsilon}$.

Proof See Appendix J. ■

Appendix G. Estimating the Hermite Moment Polynomials

In this section, we show how to estimate the Hermite moment polynomials of a regular-form mixture of Gaussians $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ where $G_j = N(\mu_j, I + \Sigma_j)$ to optimal accuracy. The main theorem that we will prove is as follows.

Theorem 65 *Consider a mixture of Gaussians $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ in $\mathbb{R}^d$. Let m be a positive integer that is sufficiently large in terms of k. Assume that $\mathcal{M}$ is in (α, β, γ)-regular form where*

- $\alpha \leq \mathrm{poly}(\log 1/\epsilon)$
- $\beta \leq \mathrm{poly}(\log 1/\epsilon)$
- *γ is sufficiently small in terms of k and m.*

Further assume that $w_{\min} \geq \theta$ for some constant θ. Let $n > \mathrm{poly}_{k,m}(d/\epsilon)$ for some sufficiently large polynomial. Assume that we are given an ϵ-corrupted set of n samples from $\mathcal{M}$, say $z_1, \dots, z_n$. There is an algorithm that runs in time $\mathrm{poly}_{k,m}(d/\epsilon)$ and with probability $1 - e^{-d/\epsilon}$ (over the random samples) outputs estimates $h'_1, \dots, h'_m$ for the Hermite moment polynomials of the mixture such that

$$\left\| v(h_j(X) - h'_j(X)) \right\| \leq (2 + \alpha + \beta + \theta^{-1} + \log 1/\epsilon)^{O_{k,m}(1)} \epsilon$$

for all $j \leq m$ where $h_1, \dots, h_m$ are the true Hermite moment polynomials of $\mathcal{M}$.

Recall the outline of the proof of Theorem 65 in Section 2.2.1. Assume that we have some initial estimates $\widetilde{h_1}, \dots, \widetilde{h_m}$ for the first m Hermite moment polynomials. We will first use the recurrence relations between Hermite moment polynomials to obtain estimates $\widetilde{h_{m+1}}, \dots, \widetilde{h_{2m}}$ for the first $2m$ Hermite moment polynomials. This is done in Section G.1. Next, recall that the Hermite moment polynomials $h_0, \dots, h_m$ are the means of the distributions $H_m(X, z)$ for $z \sim \mathcal{M}$. We use our estimates of $\widetilde{h_0}, \dots, \widetilde{h_{2m}}$ to compute estimates for the covariances of these distributions, say $\widetilde{\Sigma_{H_1}}, \dots, \widetilde{\Sigma_{H_m}}$. This is done in Section G.2. Using these estimates for the covariances, we can refine our estimates for the means, obtaining a finer set of estimates, say $\widehat{h_1}, \dots, \widehat{h_m}$. We prove that by iterating the above, we can get down to $\widetilde{O}(\epsilon)$ accuracy. This is done in Section G.3.

We begin with a simple consequence of the results in Section F.1, that for a mixture $\mathcal{M}$ in (α, β, γ)-regular form, we have a lower and upper bound on the covariance of $H_m(X, z)$ for $z \sim \mathcal{M}$.

Claim 30 *Let $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ where $G_j = N(\mu_j, I + \Sigma_j)$ be a mixture that is in (α, β, γ)-regular form. Let m be a positive integer. Assume that $w_{\min} \geq \theta$. Let Σ_{H_m} be the covariance of $v(H_m(X, z))$ for z drawn from $\mathcal{M}$. Then*

- $\Sigma_{H_m} \leq (2 + \alpha + \beta)^{O_m(1)} I$
- *As long as γ is sufficiently small in terms of m then*

$$\Sigma_{H_m} \geq \frac{1}{2} \theta I \,.$$

Proof For the first part, note that

$$\Sigma_{H_m} \leq \mathbb{E}_{z \sim \mathcal{M}} \left[v(H_m(X, z)) \otimes v(H_m(X, z)) \right] = \sum_{j=1}^{k} w_j \mathbb{E}_{z \sim G_j} \left[v(H_m(X, z)) \otimes v(H_m(X, z)) \right] \,.$$

Applying Claim 26 completes the proof of the first part.

Now we prove the second part. For a Gaussian G_j, let Σ_{H_m, G_j} be the covariance of $v(H_m(X, z))$ for z drawn from G_j. Note that

$$\Sigma_{H_m} \geq w_j \Sigma_{H_m, G_j} \geq \theta \Sigma_{H_m, G_j}$$

for all j. Now since the original mixture is in (α, β, γ)-regular form, taking j to be the component such that $\|\mu_j\| + \|\Sigma_j\|_2 \leq \gamma$ and applying Claim 27 completes the proof. ■

G.1. More Recurrence Relations

The main goal in this subsection is to prove that given ϵ-accurate estimates of $h_0, \dots, h_m$ where m is sufficiently large in terms of k, then we can compute $\widetilde{O}(\epsilon)$-accurate estimates for $h_{m+1}, \dots, h_{2m}$.

The subroutine will rely heavily on recurrence relations between the Hermite moment polynomials. Recall Lemma 52. Applying Lemma 52 to a mixture of Gaussians $\mathcal{M}$ implies that the Hermite moment polynomials $h_{j,\mathcal{M}}$ satisfy a recurrence of order $O_k(1)$ whose coefficients are "simple". We will now develop additional tools based on these recurrence relations that will allow us to estimate $h_{m+1}, \dots, h_{2m}$ using $h_0, \dots, h_m$.

In this section, we use the following notation.

- We have a mixture of k Gaussians $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ where $G_j = N(\mu_j, I + \Sigma_j)$.
- Recall Corollary 40. It will be particularly important to consider the generating function
$$f_{\mathcal{M}}(X, y) = \sum_{m=0}^{\infty} \frac{1}{m!} \cdot h_{m,\mathcal{M}}(X) y^m = w_1 e^{\mu_1(X)y + \frac{1}{2}\Sigma_1(X)y^2} + \cdots + w_k e^{\mu_k(X)y + \frac{1}{2}\Sigma_k(X)y^2} .$$
- Similar to Section E, we use $\mathcal{D}_j$ to denote the differential operator $(\partial - (\mu_j(X) + \Sigma_j(X)y))$ where the partial derivative is taken with respect to y.

The next result builds on Lemma 52 and says that for *any* recurrrence of order $2\kappa = 2(2^k - 1)$ that the first $O_k(1)$ Hermite moment polynomials almost satisfy, we can *extend* the recurrence to estimate the next several Hermite moment polynomials.

Claim 31 *Let $\kappa = 2^k - 1$. Let $T_{j,l}(X)$ for $0 \leq j \leq \kappa - 1$ and $0 \leq l \leq \kappa - j$ be polynomials such that $T_{j,l}(X)$ is homogeneous in X of degree $\kappa - j + l$.*

Let $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ be a mixture of k Gaussians in (α, β)-regular form. Let m be a positive integer that is sufficiently large in terms of k. For all $a \geq \kappa$, let
$$D_a(X) = h_{a,\mathcal{M}}(X) + (a - \kappa)! \sum_{j=0}^{\kappa-1} \sum_{l=0}^{\kappa-j} \frac{h_{a-\kappa+j-l,\mathcal{M}}(X) T_{j,l}(X)}{(a - \kappa - l)!} ,$$
where undefined terms (i.e. negative factorials in the denominator) are treated as 0. Assume that for all $a \leq m$,
$$\|v(D_a(X))\| \leq \epsilon .$$
Then for all $m \leq a \leq 2m$,
$$\|v(D_a(X))\| \leq (2 + \alpha)^{O_{m,k}(1)} \epsilon .$$

Proof

For each j with $0 \leq j \leq \kappa - 1$, let
$$T_j(X, y) = \sum_{l=0}^{\kappa-j} T_{j,l}(X) y^l .$$

Now define the differential operator

$$\mathcal{T} = \partial^{\kappa} + T_{\kappa-1}(X,y)\partial^{\kappa-1} + \cdots + T_0(X,y)\,.$$

Consider the generating function $\mathcal{T}(f_{\mathcal{M}}(X,y))$. Note that its power series expansion is precisely

$$\mathcal{T}(f_{\mathcal{M}}(X,y)) = \sum_{j=0}^{\infty} \frac{D_{\kappa+j}(X)y^j}{j!}\,.$$

Alternatively, applying the operator $\mathcal{T}$ to the sum-of-exponentials form of $f_{\mathcal{M}}$, we see that $\mathcal{T}(f_{\mathcal{M}}(X,y))$ can be written in the form

$$\mathcal{T}(f_{\mathcal{M}}(X,y)) = P_1(X,y)e^{\mu_1(X)y+\frac{1}{2}\Sigma_1(X)y^2} + \cdots + P_k(X,y)e^{\mu_k(X)y+\frac{1}{2}\Sigma_k(X)y^2}$$

where each of the polynomials $P_1,\dots,P_k$ has degree at most κ in y. Thus if we let

$$\mathcal{D} = \mathcal{D}_k^{2^{k-1}(\kappa+1)}\mathcal{D}_{k-1}^{2^{k-2}(\kappa+1)}\dots\mathcal{D}_1^{(\kappa+1)}$$

then by repeatedly applying Claim 21, we get

$$\mathcal{D}\left(\mathcal{T}(f_{\mathcal{M}}(X,y))\right) = 0\,.$$

Note that we can use Claim 18 to write the differential operator $\mathcal{D}$ in the form

$$\mathcal{D} = \partial^{\kappa(\kappa+1)} + R_{\kappa(\kappa+1)-1}(X,y)\partial^{\kappa(\kappa+1)-1} + \cdots + R_0(X,y)\,.$$

We can then write each R_j in the form

$$R_j(X,y) = R_{j,\kappa(\kappa+1)-j}(X)y^{\kappa(\kappa+1)-j} + \cdots + R_{j,0}(X)$$

where each of the polynomials $R_{j,l}$ is homogeneous in X with degree $\kappa(\kappa+1) - j + l$ and is $(O_k(1), O_k(1))$-simple with respect to $\{\mu_1(X), \Sigma_1(X), \dots, \mu_k(X), \Sigma_k(X)\}$. Using that

$$\mathcal{D}\left(\mathcal{T}(f_{\mathcal{M}}(X,y))\right) = 0\,.$$

we can now write a recurrence relation that the polynomials $D_a(X)$ must satisfy. In particular, if we let $\lambda = \kappa(\kappa+1)$, we must have for all $a \geq \lambda$

$$\frac{D_{\kappa+a}(X)}{(a-\lambda)!} + \sum_{j=0}^{\lambda-1}\sum_{l=0}^{\lambda-j} \frac{D_{\kappa+a-\lambda+j-l}(X)R_{j,l}(X)}{(a-\lambda-l)!} = 0\,.$$

This rearranges into

$$D_{\kappa+a}(X) = -(a-\lambda)!\sum_{j=0}^{\lambda-1}\sum_{l=0}^{\lambda-j} \frac{D_{\kappa+a-\lambda+j-l,}(X)R_{j,l}(X)}{(a-\lambda-l)!}\,. \tag{9}$$

Now we can use the recurrence (9) to compute $D_{m+1}(X),\dots,D_{2m}(X)$ in terms of the earlier terms in the sequence. Note that we are given

$$\|v(D_a(X))\| \leq \epsilon$$

for all $a \leq m$. Also, since $R_{j,l}$ is $(O_k(1), O_k(1))$-simple with respect to $\{\mu_1(X), \Sigma_1(X), \dots, \mu_k(X), \Sigma_k(X)\}$, we have by Claim 5 that

$$\|v(R_{j,l}(X))\| \leq (2+\alpha)^{O_k(1)} .$$

Thus, when applying the recurrence to compute $D_{m+1}(X), \dots, D_{2m}(X)$, we have for all $a \leq 2m$,

$$\|v(D_a(X))\| \leq (2+\alpha)^{O_{m,k}(1)}\epsilon .$$

■

As a consequence of the above, given estimates for the Hermite moment polynomials $h_0, \dots, h_m$, we can estimate the next Hermite moment polynomials $h_{m+1}, \dots, h_{2m}$ by first solving for a recurrence relation that the first m Hermite moment polynomials satisfy, and then extending this recurrence to compute $h_{m+1}, \dots, h_{2m}$.

Claim 32 *Let $\mathcal{M} = w_1G_1 + \dots + w_kG_k$ be a mixture of k Gaussians in (α, β)-regular form. Let m be a positive integer that is sufficiently large in terms of k. Assume that we are given estimates $h'_0(X), \dots, h'_m(X)$ such that*

$$\left\|v(h_a(X) - h'_a(X))\right\| \leq \epsilon$$

for all $a \leq m$ (where $h_a(X)$ are the true Hermite moment polynomials of $\mathcal{M}$). Then there is an algorithm that runs in $\mathrm{poly}_{m,k}(d)$ time that computes estimates $h'_{m+1}(X), \dots h'_{2m}(X)$ such that for all $a \leq 2m$

$$\left\|v(h_a(X) - h'_a(X))\right\| \leq (2+\alpha)^{O_{m,k}(1)}\epsilon .$$

Proof Let $\kappa = 2^k - 1$. We first solve for polynomials $R'_{j,l}(X)$ for all $0 \leq j \leq \kappa$ and $0 \leq l \leq \kappa - j$ such that

- $R'_{j,l}$ is homogeneous of degree $\kappa - j + l$ and $R'_{\kappa,0} = 1$
- $\left\|v(R'_{j,l}(X))\right\| \leq (2+\alpha)^{O_{m,k}(1)}$
- For all $a \leq m$

$$\left\|v\left(\sum_{j=0}^{\kappa}\sum_{l=0}^{\kappa-j}\frac{h'_{a-\kappa+j-l}(X)R'_{j,l}(X)}{(a-\kappa-l)!}\right)\right\| \leq (2+\alpha)^{O_{m,k}(1)}\epsilon$$

Note that the expressions inside the norms on the LHS are linear in the coefficients of the $R'_{j,l}$. Thus, we can solve for the coefficients via a convex program. To see that a solution exists, let $R'_{j,l}(X) = R_{j,l}(X)$ where the $R_{j,l}$ are the polynomials given by Lemma 52. It is clear that the first two conditions are satisfied because the $R_{j,l}$ are $(O_k(1), O_k(1))$-simple with respect to $\{\mu_j(X)\}_{j\in[k]}, \{\Sigma_j(X)\}_{j\in[k]}$. Next,

$$\left\|v\left(\sum_{j=0}^{\kappa}\sum_{l=0}^{\kappa-j}\frac{h'_{a-\kappa+j-l}(X)R_{j,l}(X)}{(a-\kappa-l)!}\right)\right\| = \left\|v\left(\sum_{j=0}^{\kappa}\sum_{l=0}^{\kappa-j}\frac{(h'_{a-\kappa+j-l}(X) - h_{a-\kappa+j-l}(X))R_{j,l}(X)}{(a-\kappa-l)!}\right)\right\|$$
$$\leq (2+\alpha)^{O_{m,k}(1)}\epsilon$$

for all $a \leq m$, where the last step holds because $\|v(h_a(X) - h'_a(X))\| \leq \epsilon$ for all $a \leq m$ and $\|v(R_{j,l}(X))\| \leq (2+\alpha)^{O_{m,k}(1)}$ by Claim 5.

Now we consider what happens when we apply the recurrence given by the $R'_{j,l}$, namely

$$h'_a(X) = -(a-\kappa)! \sum_{j=0}^{\kappa-1} \sum_{l=0}^{\kappa-j} \frac{h'_{a-\kappa+j-l}(X) R'_{j,l}(X)}{(a-\kappa-l)!} \tag{10}$$

and use the above to compute estimates $h'_a(X)$ for $m+1 \leq a \leq 2m$. Note that for $a \leq m$,

$$\begin{aligned}\left\| v\left(\sum_{j=0}^{\kappa} \sum_{l=0}^{\kappa-j} \frac{h_{a-\kappa+j-l}(X) R'_{j,l}(X)}{(a-\kappa-l)!} \right) \right\| &\leq \left\| v\left(\sum_{j=0}^{\kappa} \sum_{l=0}^{\kappa-j} \frac{(h_{a-\kappa+j-l}(X) - h'_{a-\kappa+j-l}(X)) R'_{j,l}(X)}{(a-\kappa-l)!} \right) \right\| \\ &\quad + \left\| v\left(\sum_{j=0}^{\kappa} \sum_{l=0}^{\kappa-j} \frac{h'_{a-\kappa+j-l}(X) R'_{j,l}(X)}{(a-\kappa-l)!} \right) \right\| \\ &\leq (2+\alpha)^{O_{m,k}(1)} \epsilon \,.\end{aligned}$$

By Claim 31, we deduce that

$$\left\| v\left(\sum_{j=0}^{\kappa} \sum_{l=0}^{\kappa-j} \frac{h_{a-\kappa+j-l}(X) R'_{j,l}(X)}{(a-\kappa-l)!} \right) \right\| \leq (2+\alpha)^{O_{m,k}(1)} \epsilon$$

for all $a \leq 2m$. Finally, comparing to the recurrence in (10) and subtracting, we get that for all $m+1 \leq a \leq 2m$,

$$\begin{aligned}\left\| v(h'_a(X) - h_a(X)) \right\| &\leq (a-\kappa)! \left\| v\left(\sum_{j=0}^{\kappa-1} \sum_{l=0}^{\kappa-j} \frac{(h'_{a-\kappa+j-l}(X) - h_{a-\kappa+j-l}(X)) R'_{j,l}(X)}{(a-\kappa-l)!} \right) \right\| \\ &\quad + (a-\kappa)! \left\| v\left(\sum_{j=0}^{\kappa} \sum_{l=0}^{\kappa-j} \frac{h_{a-\kappa+j-l}(X) R'_{j,l}(X)}{(a-\kappa-l)!} \right) \right\| \\ &\leq (2+\alpha)^{O_{m,k}(1)} (\epsilon + \max_{b<a} \left\| v(h'_b(X) - h_b(X)) \right\|) \,.\end{aligned}$$

Since we need to apply the recurrence at most m times to get the terms $h'_a(X)$ for $a \leq 2m$, the above implies that for all $a \leq 2m$

$$\left\| v(h_a(X) - h'_a(X)) \right\| \leq (2+\alpha)^{O_{m,k}(1)} \epsilon$$

as desired. It is clear that all of the steps, solving for the $R'_{j,l}$ and computing subsequent terms using the recurrence, can be done in $\mathrm{poly}_{k,m}(d)$ time. ■

G.2. Computing the Covariance of the Hermite Moment Polynomials

In the previous section, we showed how to compute $h_{m+1}, \dots, h_{2m}$ from $h_0, \dots, h_m$. Now we show how to express the covariances of the distributions of $H_0(X, z), \dots, H_m(X, z)$ (for $z \sim \mathcal{M}$) in terms of $h_0, \dots, h_{2m}$.

Claim 33 *Let $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ be a mixture of k Gaussians in (α, β)-regular form. Let Σ_{H_m} be the covariance of $v(H_m(X, z))$ for z drawn from $\mathcal{M}$. Assume that we are given estimates, say $h'_0(X), \dots, h'_{2m}(X)$, such that*

$$\left\| v(h_j(X) - h'_j(X)) \right\| \leq \epsilon$$

for all $j \leq 2m$ where $h_j(X)$ are the Hermite moment polynomials of the mixture $\mathcal{M}$. Then in $\mathrm{poly}_{k,m}(d)$ *time, we can compute a symmetric matrix M' such that*

$$\left\| M' - \Sigma_{H_m} \right\|_{op} \leq (2 + \alpha)^{O_m(1)} \epsilon \,.$$

Proof Consider two sets of d variables $X^{(1)} = (X_1^{(1)}, \dots, X_d^{(1)})$ and $X^{(2)} = (X_1^{(2)}, \dots, X_d^{(2)})$. Let

$$P(X^{(1)}, X^{(2)}) = \mathbb{E}_{z \sim \mathcal{M}} \left[H_m(X^{(1)}, z) H_m(X^{(2)}, z) \right] .$$

We will first show how to estimate the polynomial P. Define

$$\mathcal{A}(y) = \sum_{j=1}^{k} w_j e^{y \mu_j (X^{(1)} + X^{(2)}) + \frac{1}{2} y^2 \Sigma_j (X^{(1)} + X^{(2)})} .$$

By Corollary 40,

$$\mathcal{A}(y) = \sum_{j=0}^{\infty} \frac{h_j (X^{(1)} + X^{(2)}) y^j}{j!} .$$

On the other hand, let $Q_0, Q_1, \dots, Q_j$ be the terms of the power series

$$e^{y^2 (X^{(1)} \cdot X^{(2)})} \mathcal{A}(y) = \sum_{j=1}^{k} w_j e^{y \mu_j (X^{(1)} + X^{(2)}) + \frac{1}{2} y^2 \Sigma_j (X^{(1)} + X^{(2)}) + y^2 (X^{(1)} \cdot X^{(2)})} = \sum_{j=0}^{\infty} \frac{Q_j (X^{(1)}, X^{(2)}) y^j}{j!} . \tag{11}$$

Let $Q_{m,m}$ be the degree (m, m) part of Q_{2m}. By Claim 25, we know that

$$P(X^{(1)}, X^{(2)}) = \binom{2m}{m}^{-1} Q_{m,m} (X^{(1)}, X^{(2)}) . \tag{12}$$

In particular, we can use our estimates $h'_0, \dots h'_{2m}$ to estimate the first $2m + 1$ terms of $\mathcal{A}$ and then multiply by

$$e^{y^2 (X^{(1)} \cdot X^{(2)})} = \sum_{j=0}^{\infty} \frac{(X^{(1)} \cdot X^{(2)})^j y^{2j}}{j!}$$

which is an explicit power series that we can compute. Formally, let $h_{j,j}(X^{(1)}, X^{(2)})$ denote the degree j, j part of $h_{2j}(X^{(1)} + X^{(2)})$. Similarly, we use $h'_{j,j}$ to denote the degree j, j part of $h'_{2j}(X^{(1)} + X^{(2)})$. Using (11)

$$\sum_{j=0}^{\infty} \frac{Q_j(X^{(1)}, X^{(2)})y^j}{j!} = \left(\sum_{j=0}^{\infty} \frac{(X^{(1)} \cdot X^{(2)})^j y^{2j}}{j!}\right) \left(\sum_{j=0}^{\infty} \frac{h_j(X^{(1)} + X^{(2)})y^j}{j!}\right) .$$

Thus, by (12),

$$P(X^{(1)}, X^{(2)}) = (m!)^2 \sum_{j=0}^{m} \frac{h_{j,j}(X^{(1)} + X^{(2)})(X^{(1)} \cdot X^{(2)})^{m-j}}{(2j)!(m-j)!} .$$

On the other hand, using our estimates h', we may compute

$$P'(X^{(1)}, X^{(2)}) = (m!)^2 \sum_{j=0}^{m} \frac{h'_{j,j}(X^{(1)} + X^{(2)})(X^{(1)} \cdot X^{(2)})^{m-j}}{(2j)!(m-j)!} .$$

Note that since

$$\left\| v(h_j(X) - h'_j(X)) \right\| \leq \epsilon$$

for all $j \leq 2m$, we have that

$$\left\| v(h_j(X^{(1)} + X^{(2)}) - h'_j(X^{(1)} + X^{(2)})) \right\| \leq O_m(1)\epsilon$$

for all $j \leq 2m$ where for the vectorization we view the polynomial as a homogeneous polynomial in $2d$ variables. The above implies

$$\left\| v(h_{j,j}(X^{(1)} + X^{(2)}) - h'_{j,j}(X^{(1)} + X^{(2)})) \right\| \leq O_m(1)\epsilon$$

for all $j \leq m$ so we conclude

$$\left\| T_{\mathsf{sym}}(h'_{j,j}(X^{(1)} + X^{(2)}) - h_{j,j}(X^{(1)} + X^{(2)})) \right\|_2 \leq O_m(1)\epsilon$$

for all $j \leq m$. Now since

$$T_{\mathsf{sym}}\left((X^{(1)} \cdot X^{(2)})^{m-j}\right) = I$$

we can use Claim 10 to deduce

$$\left\| T_{\mathsf{sym}}\left(P(X^{(1)}, X^{(2)}) - P'(X^{(1)}, X^{(2)})\right) \right\|_{\mathsf{op}} \leq O_m(1)\epsilon \, . \tag{13}$$

However, note that by Claim 9

$$T_{\mathsf{sym}}\left(P(X^{(1)}, X^{(2)})\right) = \mathbb{E}_{z \sim \mathcal{M}}\left[v(H_m(X, z)) \otimes v(H_m(X, z))\right] . \tag{14}$$

To compute Σ_{H_m}, it remains to estimate

$$E_{z \sim \mathcal{M}}[v(H_m(X, z))] \otimes E_{z \sim \mathcal{M}}[v(H_m(X, z))] = v(h_m(X)) \otimes v(h_m(X)) \, .$$

We can simply use our estimates $h'_m(X)$ and compute $v(h'_m(X)) \otimes v(h'_m(X))$. Note

$$\begin{aligned}
&\left\|v(h_m(X)) \otimes v(h_m(X)) - v(h'_m(X)) \otimes v(h'_m(X))\right\|_{\mathsf{op}} \\
&\leq \left\|v(h'_m(X)) \otimes v(h'_m(X) - h_m(X))\right\|_{\mathsf{op}} + \left\|v(h'_m(X) - h_m(X)) \otimes v(h_m(X))\right\|_{\mathsf{op}} \\
&\leq \left\|v(h'_m(X) - h_m(X))\right\|_2 \left(\left\|v(h'_m(X)\right\|_2 + \|v(h_m(X)\|_2\right) .
\end{aligned}$$

However using Claim 26, we know that $\|v(h_m(X)\| \leq (2+\alpha)^{O_m(1)}$. Thus, the RHS of the above is at most $(2+\alpha)^{O_m(1)}\epsilon$.

Overall, we can compute

$$M' = T_{\mathsf{sym}}(P'(X^{(1)}, X^{(2)})) - v(h'_m(X)) \otimes v(h'_m(X))$$

and combining everything we have shown so far (13, 14) gives

$$\begin{aligned}
\left\|M' - \Sigma_{H_m}\right\|_{\mathsf{op}} \leq &\left\|T_{\mathsf{sym}}(P'(X^{(1)}, X^{(2)})) - \mathbb{E}_{z\sim\mathcal{M}}\left[v(H_m(X,z)) \otimes v(H_m(X,z))\right]\right\|_{\mathsf{op}} \\
&+ \left\|v(h'_m(X)) \otimes v(h'_m(X)) - E_{z\sim\mathcal{M}}[v(H_m(X,z))] \otimes E_{z\sim\mathcal{M}}[v(H_m(X,z))]\right\|_{\mathsf{op}} \\
= &\left\|T_{\mathsf{sym}}\left(P'(X^{(1)}, X^{(2)}) - P(X^{(1)}, X^{(2)})\right)\right\|_{\mathsf{op}} \\
&+ \left\|v(h_m(X)) \otimes v(h_m(X)) - v(h'_m(X)) \otimes v(h'_m(X))\right\|_{\mathsf{op}} \\
\leq &\,(2+\alpha)^{O_m(1)}\epsilon
\end{aligned}$$

It is clear that computing M' can be done in $\mathrm{poly}_{k,m}(d)$ time so we are done. ■

G.3. Estimating the Hermite Moment Polynomials Optimally

We can now put together all of the parts in the previous two subsections to get an algorithm for estimating the Hermite moment polynomials to within $\widetilde{O}(\epsilon)$ accuracy, proving Theorem 65.

We will rely on the following generic theorems about robustly estimating the mean of a distribution from Diakonikolas and Kane (2019).

Theorem 66 (Theorem 2.7 in Diakonikolas and Kane (2019)) *Let S be a $(3\epsilon, \delta)$-stable set with respect to a distribution X and let T be an ϵ-corrupted version of S. There is a polynomial time algorithm which given T returns $\widehat{\mu}$ such that*

$$\|\widehat{\mu} - \mu_X\| = O(\delta) .$$

Corollary 67 (Corollary 2.9 in Diakonikolas and Kane (2019)) *Let T be an ϵ-corrupted set of samples of size at least $\mathrm{poly}(d/\epsilon)$ for some sufficiently large polynomial from a distribution X on $\mathbb{R}^d$ with unknown bounded covariance $\Sigma_X \leq \sigma^2 I$. There exists a polynomial time algorithm which given T returns $\widehat{\mu}$ such that with $1 - e^{-d/\epsilon}$ probability*

$$\|\widehat{\mu} - \mu_X\| = O(\sigma\sqrt{\epsilon}) .$$

Theorem 65 will follow immediately from the next lemma. The lemma states that given estimates for the first m Hermite moment polynomials that are accurate to within some parameter η, we can refine them to be accurate to within roughly $\sqrt{\epsilon\eta}$. The proof of the lemma involves combining the results in Sections G.1 and G.2 with Theorem 66. To see how Theorem 65 follows from the lemma, we can use Corollary 67 to obtain initial estimates and then iterate Lemma 68.

Lemma 68 *Consider a mixture of Gaussians $\mathcal{M} = w_1G_1 + \cdots + w_kG_k$ in $\mathbb{R}^d$. Let m be a positive integer that is sufficiently large in terms of k. Assume that $\mathcal{M}$ is in (α, β, γ)-regular form where*

- $\alpha \leq \text{poly}(\log 1/\epsilon)$
- $\beta \leq \text{poly}(\log 1/\epsilon)$
- *γ is sufficiently small in terms of k and m.*

Further assume that $w_{\min} \geq \theta$ for some constant θ. Let $n > \text{poly}_{k,m}(d/\epsilon)$ for some sufficiently large polynomial. Assume that we are given an ϵ-corrupted set of n samples from $\mathcal{M}$, say $z_1, \dots, z_n$. Also assume that we are given estimates $\widetilde{h_1}, \dots, \widetilde{h_m}$ of the Hermite moment polynomials such that

$$\left\| v(h_j(X) - \widetilde{h_j}(X)) \right\| \leq \eta$$

for all $j \leq m$ where η is a parameter such that $\eta \leq 1/\text{poly}_{k,m}(\alpha, \beta, \theta^{-1}, \log 1/\epsilon)$. Then there is an algorithm that runs in $\text{poly}_{k,m}(d/\epsilon)$ time and, with probability at least $1 - e^{-10d/\epsilon}$ over the random samples, outputs estimates $\widehat{h_1}, \dots, \widehat{h_m}$ such that

$$\left\| v(h_j(X) - \widehat{h_j}(X)) \right\| \leq (2 + \alpha + \beta + \theta^{-1} + \log 1/\epsilon)^{O_{m,k}(1)} \sqrt{\epsilon\eta}\,.$$

Proof By Claim 32 and Claim 33, we can compute estimates $\Sigma^0_{H_1}, \dots, \Sigma^0_{H_m}$ such that for all $j \leq m$

$$\left\| \Sigma^0_{H_j} - \Sigma_{H_j} \right\|_{\mathsf{op}} \leq (2 + \alpha + \beta)^{O_{k,m}(1)} \eta \tag{15}$$

where Σ_{H_j} is the covariance of $v(H_j(X, z))$ for z drawn from $\mathcal{M}$. Now by Claim 30, each of these estimates must be positive semi-definite and have $\Sigma^0_{H_j} \geq \Omega(\theta) I$ so we can take their positive semidefinite square roots. We have

$$\left\| I - (\Sigma^0_{H_j})^{-1/2} \Sigma_{H_j} (\Sigma^0_{H_j})^{-1/2} \right\|_{\mathsf{op}} \leq (2 + \alpha + \beta + \theta^{-1})^{O_{k,m}(1)} \eta\,. \tag{16}$$

Now fix a $j \leq m$. Let $\mathcal{D}_j$ be the distribution of $v(H_j(X, z))$ for $z \sim \mathcal{M}$. The above implies that the covariance of the distribution obtained after applying $(\Sigma^0_{H_j})^{-1/2}$ to $\mathcal{D}_j$ is close to identity.

By Claim 29, with $1 - e^{-10d/\epsilon}$ probability, a set of n uncorrupted samples is $(3\epsilon, \delta)$-pseudo-stable with respect to $\mathcal{D}_j$ with

$$\delta = (2 + \alpha + \beta + \log 1/\epsilon)^{O_{m,k}(1)} \epsilon\,.$$

If this holds, then combining the pseudo-stability with (16) and the fact that $\Sigma^0_{H_j} \geq \Omega(\theta) I$ implies that the set of uncorrupted samples obtained after applying the transformation $(\Sigma^0_{H_j})^{-1/2}$ is stable with respect to the distribution $(\Sigma^0_{H_j})^{-1/2} \mathcal{D}_j$ with parameters

$$\left(3\epsilon, (2 + \alpha + \beta + \theta^{-1} + \log 1/\epsilon)^{O_{m,k}(1)} \sqrt{\epsilon\eta}\right)\,.$$

Now by Theorem 66, we can estimate the mean of $(\Sigma^0_{H_j})^{-1/2}\mathcal{D}_j$ up to accuracy

$$(2+\alpha+\beta+\theta^{-1}+\log 1/\epsilon)^{O_{m,k}(1)}\sqrt{\epsilon\eta}\,.$$

Claim 30 and (15) imply that the operator norm of $(\Sigma^0_{H_j})^{1/2}$ is bounded by $(2+\alpha+\beta)^{O_{m,k}(1)}$ so now we can simply invert the linear transformation and estimate the mean of $\mathcal{D}_j$ to within

$$(2+\alpha+\beta+\theta^{-1}+\log 1/\epsilon)^{O_{m,k}(1)}\sqrt{\epsilon\eta}\,.$$

However, the mean of $\mathcal{D}_j$ is exactly $v(h_j(X))$ so repeating this for all j completes the proof. ■

We now prove Theorem 65 by iterating Lemma 68.

Proof [Proof of Theorem 65]

Note that $v(h_j(X))$ is the mean of $v(H_m(X,z))$ for z drawn from $\mathcal{M}$. Now by Claim 30 and Corollary 67, we can obtain estimates

$$\widetilde{h_1}(X),\dots,\widetilde{h_m}(X)$$

of the Hermite moment polynomials such that for all $j \leq m$

$$\left\|v(h_j(X)-\widetilde{h_j}(X))\right\| \leq (2+\alpha+\beta)^{O_{k,m}(1)}\sqrt{\epsilon}\,.$$

Now we can iterate Lemma 68 on these estimates until we obtain final estimates $h_1',\dots,h_m'$ such that

$$\left\|v(h_j(X)-h_j'(X))\right\| \leq (2+\alpha+\beta+\theta^{-1}+\log 1/\epsilon)^{O_{k,m}(1)}\epsilon$$

for all $j \leq m$ (to see this, note that each time the above inequality is not true, when we apply Lemma 68 to refine our estimates, we reduce our estimation error by a factor of $1/2$). ■

Appendix H. Learning Regular Form Mixtures

In this section, we combine everything that we have shown so far to give an algorithm for learning regular-form mixtures of Gaussians. Recall the outline in Section 2.2.1. We have some unknown mixture $\mathcal{M} = w_1G_1+\dots+w_kG_k$ in regular form. In this section, we will assume that we are given estimates $\overline{G_1},\dots,\overline{G_k}$ for the components such that

$$d_{\mathsf{TV}}(G_j,\overline{G_j}) \leq \epsilon^c$$

for some constant $c>0$. These estimates can be obtained by directly applying results from Liu and Moitra (2021) (see Section I.1 for more details).

We will then bootstrap the rough component estimates by multiplying appropriate polynomials in front of them. In particular, we show how to compute a degree-$O_{k,c}(1)$ MPG distribution

$$f = Q_1(x)\overline{G_1}(x)+\dots+Q_k(x)\overline{G_k}(x)$$

such that

$$\|\mathcal{M}-f\|_1 \leq \widetilde{O}(\epsilon)\,.$$

The main theorem that we will prove in this section is stated formally below.

Theorem 69 *Let $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ where $G_j = N(\mu_j, I + \Sigma_j)$ be a mixture of Gaussians such that all mixing weights are at least θ where $\theta^{-1} \leq \mathrm{poly}(\log 1/\epsilon)$. Let $c > 0$ be a (small) constant. Assume that the mixture is in (α, β, γ)-regular form where*

- $\alpha = \mathrm{poly}(\log 1/\epsilon)$
- $\beta = \mathrm{poly}(\log 1/\epsilon)$
- *γ is sufficiently small in terms of k and c.*

Let $n > \mathrm{poly}_{k,c}(d/\epsilon)$ for some sufficiently large polynomial. Assume that we are given an ϵ-corrupted set of n samples $X_1, \dots, X_n$ from $\mathcal{M}$. Assume that an adversary also gives us estimates $\overline{G_1} = N(\widetilde{\mu}_1, I + \widetilde{\Sigma}_1), \dots, \overline{G_k} = N(\widetilde{\mu}_k, I + \widetilde{\Sigma}_k)$ for the components with the promise that

$$d_{\mathsf{TV}}(G_j, \overline{G_j}) \leq \epsilon^c \,.$$

There is an algorithm that runs in time $\mathrm{poly}_{k,c}(d/\epsilon)$ and with high probability (over the random samples) outputs a degree-$O_{k,c}(1)$ MPG distribution $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = Q_1(x)\overline{G}_1(x) + \dots + Q_k(x)\overline{G}_k(x)$$

such that f satisfies

$$\|\mathcal{M}(x) - f(x)\|_1 \leq (2 + \alpha + \beta + \theta^{-1} + \log 1/\epsilon)^{O_{k,c}(1)} \epsilon \,.$$

To prove Theorem 69, we rely on the two main results that we have shown so far, Theorem 65 and Theorem 50, which are (informally) that

- We can estimate low-degree Hermite moment polynomials of $\mathcal{M}$ to nearly optimal accuracy
- If two low-degree MPG functions are close on their low-degree Hermite moment polynomials then they are close in L^1 norm

To complete the learning algorithm, it remains to actually compute a low-degree MPG distribution that matches the Hermite moment polynomial estimates obtained from Theorem 65. Then, Theorem 50 will imply that this MPG distribution is close to the density function of $\mathcal{M}$ in L^1 norm.

H.1. Preliminary Computations

We will first need a few preliminary definitions and computations.

Definition 70 *We say a Gaussian G is χ-balanced if its covariance matrix Σ satisfies*

$$\frac{1}{\chi} I \leq \Sigma \leq \chi I \,.$$

If we have a mixture $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ such that all components are χ-balanced, then we say that the mixture is χ-balanced.

Note that the parameter β governs the balancedness of a mixture in (α, β, γ)-regular form. We will need the following few basic facts.

Claim 34 *For two Gaussians $N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)$*

$$d_{\mathsf{TV}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) = O\left(\left((\mu_1 - \mu_2)^T \Sigma_1^{-1} (\mu_1 - \mu_2) \right)^{1/2} + \left\| \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} - I \right\|_F \right)$$

Proof See e.g. Fact 2.1 in Kane (2021). ■

The following two claims relate TV distance and parameter distance for balanced Gaussians. The first deals with the case when the two Gaussians have very small overlap while the second deals with the case when the two Gaussians have very large overlap.

Claim 35 *Let $G_1 = N(\mu_1, \Sigma_1), G_2 = N(\mu_2, \Sigma_2)$ be χ-balanced Gaussians. Let $\epsilon < 0.1$ be a parameter and assume that $d_{\mathsf{TV}}(G_1, G_2) \leq 1 - \epsilon$. Then the following two conditions hold:*

1. $\|\mu_1 - \mu_2\| \leq O(\sqrt{\chi \log 1/\epsilon})$
2. $\|\Sigma_1 - \Sigma_2\|_2 \leq O(\chi \log^2 1/\epsilon)$

Proof For the first claim, note that if we project onto the line connecting μ_1 and μ_2 then both distributions are Gaussian with standard deviation at most $\sqrt{\chi}$. Thus, their means must be separated by at most $O(\sqrt{\chi \log 1/\epsilon})$. For the second claim, we can use Lemma 3.2 in Kane (2021). Note that the covariance of the mixture $(G_1 + G_2)/2$ is

$$\Sigma = \frac{\Sigma_1 + \Sigma_2}{2} + \frac{(\mu_1 - \mu_2)}{2} \frac{(\mu_1 - \mu_2)^T}{2}$$

and the first part implies that

$$\Sigma \leq O(\chi \log 1/\epsilon) I \,.$$

Lemma 3.2 from Kane (2021) now immediately gives the desired bound. ■

Claim 36 *Let $G_1 = N(\mu_1, \Sigma_1), G_2 = N(\mu_2, \Sigma_2)$ be χ-balanced Gaussians. Let $\epsilon < 0.1$ be a parameter and assume that $d_{\mathsf{TV}}(G_1, G_2) \leq \epsilon$. Then the following two conditions hold:*

1. $\|\mu_1 - \mu_2\| \leq O(\sqrt{\chi}\epsilon)$
2. $\|\Sigma_1 - \Sigma_2\|_2 \leq \mathrm{poly}(\chi)\epsilon$

Proof The first part follows from Theorem 1.2 in Devroye et al. (2018). To prove the second part, let $G_1' = N(\mu_2, \Sigma_1)$. Note that by Claim 34 and the first part,

$$d_{\mathsf{TV}}(G_1', G_2) \leq d_{\mathsf{TV}}(G_1, G_2) + d_{\mathsf{TV}}(G_1, G_1') = O(\chi\epsilon) \,.$$

Now, applying Theorem 1.1 from Devroye et al. (2018), we deduce that

$$\|\Sigma_1 - \Sigma_2\|_2 \leq \mathrm{poly}(\chi)\epsilon \,,$$

completing the proof. ■

H.2. Main Proof of Theorem 69

Now we prove the first key lemma of this section. This lemma implies that there exist polynomials that we can multiply in front of our rough component estimates $\overline{G_1}, \dots, \overline{G_k}$ in order to match the Hermite moment polynomials of the true mixture $\mathcal{M}$. Afterwards, we show how to solve for these polynomials.

Lemma 71 *Let $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ be a mixture of Gaussians. Let $c > 0$ be a (small) constant. Assume that the mixture is in (α, β, γ)-regular form where*

- $\alpha = \mathrm{poly}(\log 1/\epsilon)$
- $\beta = \mathrm{poly}(\log 1/\epsilon)$
- *γ is sufficiently small in terms of k and c.*

Let $G_j = N(\mu_j, I + \Sigma_j)$ for all $j \in [k]$. Let $\overline{G_1} = N(\widetilde{\mu}_1, I + \widetilde{\Sigma}_1), \dots, \overline{G_k} = N(\widetilde{\mu}_k, I + \widetilde{\Sigma}_k)$ be Gaussians such that

$$d_{\mathsf{TV}}(G_j, \overline{G_j}) \leq \epsilon^c \,.$$

Let m be a parameter. Then there exist polynomials $P_1, \dots, P_k$ in d variables of degree at most $10/c$ such that the following holds:

1. *If we write the power series expansion of the function*

$$\sum_{j=1}^{k} w_j e^{\mu_j(X)y + \frac{1}{2}\Sigma_j(X)y^2} - \sum_{j=1}^{k} (w_j + P_j(Xy)) e^{\widetilde{\mu}_j(X)y + \frac{1}{2}\widetilde{\Sigma}_j(X)y^2} = \sum_{l=0}^{\infty} \frac{f_l(X)y^l}{l!}$$

   then

$$\|v(f_l(X))\| \leq (2 + \alpha + \beta)^{O_{k,m,c}(1)} \epsilon$$

   for all $0 \leq l \leq m$.

2. *For all j, $P_j(0) = 0$*
3. *For all j, $\|v_y(P_j(Xy))\| \leq \beta^{O_{k,m,c}(1)} \epsilon^c$*

Proof

For simplicity, assume that $10/c$ is an integer. The modification to when $10/c$ is not an integer is straight-forward. For $j \in [k]$, let

$$g^{(j)}(X, y) = e^{(\mu_j(X) - \widetilde{\mu}_j(X))y + \frac{1}{2}(\Sigma_j(X) - \widetilde{\Sigma}_j(X))y^2} = \sum_{l=0}^{\infty} \frac{g_l^{(j)}(X)y^l}{l!}$$

where for the last expression, we expand the generating function as a power series in y. Let

$$h^{(j)}(X, y) = \sum_{l=0}^{10/c} \frac{g_l^{(j)}(X)y^l}{l!}$$

i.e. $h^{(j)}$ is obtained by truncating the power series expansion of $g^{(j)}$ to the first $10/c$ terms. Note that in the RHS above, $g_l^{(j)}(X)$ is homogeneous in X of degree l. Thus, we can write $h^{(j)}(X,y)$ as a polynomial in the d-tuple of formal variables Xy. We claim that setting $P_1,\dots,P_k$ such that

$$P_j(Xy) = w_j(h^{(j)}(X,y)-1) = w_j\sum_{l=1}^{10/c}\frac{g_l^{(j)}(X)y^l}{l!}$$

for all $j\in[k]$ suffices.

Note that the setting trivially satisfies condition 2. Next, we check condition 3. First note that because $d_{\mathsf{TV}}(G_j,\overline{G_j})\le\epsilon^c$, Claim 36 implies that

$$\|v(\mu_j(X)-\widetilde{\mu}_j(X))\|\,,\left\|v(\Sigma_j(X)-\widetilde{\Sigma}_j(X))\right\|\le\beta^{O(1)}\epsilon^c\,.$$

Thus, since

$$e^{(\mu_j(X)-\widetilde{\mu}_j(X))y+\frac{1}{2}(\Sigma_j(X)-\widetilde{\Sigma}_j(X))y^2} = 1+\sum_{l=1}^{\infty}\frac{\left((\mu_j(X)-\widetilde{\mu}_j(X))y+\frac{1}{2}(\Sigma_j(X)-\widetilde{\Sigma}_j(X))y^2\right)^l}{l!}$$

we have

$$\sum_{l=1}^{\infty}\frac{g_l^{(j)}(X)y^l}{l!}=\sum_{l=1}^{\infty}\frac{\left((\mu_j(X)-\widetilde{\mu}_j(X))y+\frac{1}{2}(\Sigma_j(X)-\widetilde{\Sigma}_j(X))y^2\right)^l}{l!}\tag{17}$$

and we can use Claim 5 and the triangle inequality to verify condition 3.

Now we check condition 1. By combining (17) with Claim 5 and the triangle inequality, we get that for all l with $10/c\le l\le m$,

$$\left\|v(g_l^{(j)}(X))\right\|\le(2+\beta)^{O_m(1)}\epsilon\,.$$

Next write the power series expansion

$$e^{\widetilde{\mu}_j(X)y+\frac{1}{2}\widetilde{\Sigma}_j(X)y^2}=\sum_{l=0}^{\infty}\frac{t_l^{(j)}(X)y^l}{l!}\,.$$

Since the true mixture is in (α,β,γ)-regular form and the components $\overline{G_j}$ are ϵ^c-close to the respective true components, we get that for all l with $0\le l\le m$,

$$\left\|v(t_l^{(j)}(X))\right\|\le(2+\alpha+\beta)^{O_m(1)}\,.$$

Finally note that

$$w_j+P_j(Xy)=w_jh^{(j)}(X,y)\,.$$

Now we may write

$$\sum_{j=1}^{k} w_j e^{\mu_j(X)y+\frac{1}{2}\Sigma_j(X)y^2} - \sum_{j=1}^{k}(w_j + P_j(Xy))e^{\widetilde{\mu}_j(X)y+\frac{1}{2}\widetilde{\Sigma}_j(X)y^2}$$
$$= \sum_{j=1}^{k} w_j \left(\sum_{l=10/c+1}^{\infty} \frac{g_l^{(j)}(X)y^l}{l!} \right) e^{\widetilde{\mu}_j(X)y+\frac{1}{2}\widetilde{\Sigma}_j(X)y^2} .$$

However, since we only care about the first m terms of the power series expansion, it suffices to consider the expression

$$\sum_{j=1}^{k} w_j \left(\sum_{l=10/c+1}^{m} \frac{g_l^{(j)}(X)y^l}{l!} \right) \left(\sum_{l=0}^{m} \frac{t_l^{(j)}(X)y^l}{l!} \right) .$$

Finally using our bounds on $\left\| v(g_l^{(j)}(X)) \right\|$ and $\left\| v(t_l^{(j)}(X)) \right\|$ from above, we immediately get the desired inequality. Note that in the above we only dealt with the case when $10/c < m$. If $10/c \geq m$, then the above argument implies that we can actually choose $P_1, \dots, P_k$ so that $f_0, \dots, f_m$ are all identically 0.

■

Using Lemma 71 we can now prove Theorem 69.

Proof [Proof of Theorem 69] Let m be a constant that will be set later as a sufficiently large function depending only on k, c. Using Theorem 65, we can obtain estimates $h_1', \dots, h_m'$ for the Hermite moment polynomials of the mixture $\mathcal{M}$ such that

$$\left\| v(h_j(X) - h_j'(X) \right\| \leq (2 + \alpha + \beta + \theta^{-1} + \log 1/\epsilon)^{O_{k,m}(1)} \epsilon \tag{18}$$

for all $j \leq m$ where $h_1, \dots, h_m$ denote the true Hermite moment polynomials of the mixture. We will solve for constants $c_1, \dots, c_k$ and polynomials $P_1(X), \dots, P_k(X)$ in variables $X = (X_1, \dots, X_d)$ of degree at most $10/c$ such that the following properties hold:

1. If we write the following generating function as a formal power series

$$(c_1 + P_1(Xy))e^{\widetilde{\mu}_1(X)y+\frac{1}{2}\widetilde{\Sigma}_1(X)y^2} + \dots + (c_k + P_k(Xy))e^{\widetilde{\mu}_k(X)y+\frac{1}{2}\widetilde{\Sigma}_k(X)y^2} = \sum_{j=0}^{\infty} \frac{f_j(X)y^j}{j!}$$

   then

$$\left\| v(h_j'(X) - f_j(X)) \right\| \leq (2 + \alpha + \beta + \theta^{-1} + \log 1/\epsilon)^{O_{k,m,c}(1)} \epsilon \tag{19}$$

   for all $j \leq m$.

2. $c_j \geq \theta$ for all j

3. $P_j(0) = 0$ and $\|v(P_j(X))\| \leq \beta^{O_{k,m,c}(1)} \epsilon^c$ for all $j \in [k]$

Note that the expressions $h_j'(X) - f_j(X)$ are linear in $c_1, \dots, c_k$ and the coefficients of $P_1, \dots, P_k$ so if the system is feasible, then we can find a solution efficiently because all constraints are convex. Note that we view it as a system where the indeterminates that we are solving for are exactly $c_1, \dots, c_k$ and the coefficients of $P_1, \dots, P_k$.

To see that the system is feasible, it suffices to combine (18) with Lemma 71 and note that by Corollary 40

$$\sum_{j=1}^k w_j e^{\mu_j(X)y + \frac{1}{2}\Sigma_j(X)y^2} = \sum_{j=0}^\infty \frac{h_j(X)y^j}{j!}\,.$$

Thus, by solving the system given by (19), we can assume that we found a valid solution $c_1, \dots, c_k$ and $P_1(X), \dots, P_k(X)$. To complete the proof, we now show that from any solution to (19), we can construct an MPG distribution f that is close to the density function of the mixture.

By choosing m sufficiently large in terms of k, c, we can now apply Theorem 58 on the following generating function

$$\sum_{j=1}^k (c_j + P_j(Xy)) e^{\widetilde{\mu}_j(X)y + \frac{1}{2}\widetilde{\Sigma}_j(X)y^2} - \sum_{j=1}^k w_j e^{\mu_j(X)y + \frac{1}{2}\Sigma_j(X)y^2}\,.$$

To verify the conditions of the theorem, we can combine (18, 19) and note that the polynomials $P_1(X), \dots, P_k(X)$ have degree at most $10/c$. We deduce that if we let

$$T(X) = \sum_{j=1}^k (c_j + P_j(iX)) e^{i\widetilde{\mu}_j(X) - \frac{1}{2}\widetilde{\Sigma}_j(X)} - \sum_{j=1}^k w_j e^{i\mu_j(X) - \frac{1}{2}\Sigma_j(X)}$$

then

$$\|\chi T\|_1 \leq (2 + \alpha + \beta + \theta^{-1} + \log 1/\epsilon)^{O_{k,c}(1)}\epsilon\,.$$

However note that by Claim 14,

$$\chi\left(\sum_{j=1}^k w_j e^{i\mu_j(X) - \frac{1}{2}\Sigma_j(X)}\right)$$

is exactly the density function of $\mathcal{M}$. Thus, setting

$$f_0 = \chi\left(\sum_{j=1}^k (c_j + P_j(iX)) e^{i\widetilde{\mu}_j(X) - \frac{1}{2}\widetilde{\Sigma}_j(X)}\right)$$

achieves that

$$\|\mathcal{M}(x) - f_0(x)\|_1 \leq (2 + \alpha + \beta + \theta^{-1} + \log 1/\epsilon)^{O_{k,c}(1)}\epsilon\,.$$

Note that by applying Claim 16, we get that f_0 is a valid low-degree MPG function. However, it is not necessarily a distribution. We now show how to make minor modifications to f_0 to transform it into a distribution. Note that by Claim 16 and condition 3 in the system that we solved for the P_j, the function f_0 defined above can be written in the form

$$f_0(x) = (c_1 + R_1(x))\tilde{G}_1(x) + \cdots + (c_k + R_k(x))\tilde{G}_k(x)$$

where for all $j \in [k]$, R_j is a polynomial with real coefficients and degree at most $10/c$ and

$$\mathbb{E}_{x \sim G_j}\left[(R_j(x))^2\right] \leq (2+\beta)^{O_{k,c}(1)}\epsilon^{2c}.$$

Note that the R_j have real coefficients. Let B be an integer such that $B > 10/c$. Note that for all $x \in \mathbb{R}^d$, by the AM-GM inequality,

$$1 + \frac{R_j(x)}{c_j} + \left(\frac{R_j(x)}{c_j}\right)^{2B} \geq 0.$$

Now let

$$f_1(x) = \sum_{j=1}^{k}\left(c_j + R_j(x) + \frac{R_j(x)^{2B}}{c_j^{2B-1}}\right)\tilde{G}_j(x).$$

We verify that f_1 is close to $\mathcal{M}$ in L^1 norm. Note that using hypercontractivity (Claim 4), we have

$$\begin{aligned}\|f_1 - f_0\|_1 \leq \theta^{-(2B-1)}\sum_{j=1}^{k}\mathbb{E}_{x\sim G_j}\left[R_j(x)^{2B}\right] \leq (\theta^{-1})^{O_{k,c}(1)}\sum_{j=1}^{k}\left(\mathbb{E}_{x\sim G_j}\left[(R_j(x))^2\right]\right)^B \\ \leq (2+\beta+\theta^{-1})^{O_{k,c}(1)}\epsilon.\end{aligned}$$

Thus,

$$\|\mathcal{M}(x) - f_1(x)\|_1 \leq (2+\alpha+\beta+\theta^{-1}+\log 1/\epsilon)^{O_{k,c}(1)}\epsilon.$$

Finally, note that the above implies that

$$1-(2+\alpha+\beta+\theta^{-1}+\log 1/\epsilon)^{O_{k,c}(1)}\epsilon \leq \int_{\mathbb{R}^d} f_1(x)dx \leq 1+(2+\alpha+\beta+\theta^{-1}+\log 1/\epsilon)^{O_{k,c}(1)}\epsilon$$

and furthermore, we can explicitly compute the integral $\int_{\mathbb{R}^d} f_1(x)dx$ in polynomial time so letting

$$f = \frac{f_1(x)}{\int_{\mathbb{R}^d} f_1(x)dx}$$

yields a degree-$O_{k,c}(1)$ MPG distribution such that

$$\|\mathcal{M}(x) - f(x)\|_1 \leq (2+\alpha+\beta+\theta^{-1}+\log 1/\epsilon)^{O_{k,c}(1)}\epsilon,$$

which completes the proof. ■

Appendix I. Full Algorithm

Now we are ready to complete our full learning algorithm. A high-level description of our full algorithm is given below. Our algorithm consists of the following steps. First, we show that we can estimate the components of the mixture to $\epsilon^{\Omega(1)}$ accuracy by modifying the techniques in Liu and Moitra (2021). Then using these estimates, we cluster into submixtures such that the clustering is $\widetilde{O}(\epsilon)$-accurate. Finally, for each submixture, we argue that we can compute a linear transformation that places it in regular form and then apply Theorem 69 to estimate its density function. Since we will enumerate over many possible candidate clusterings, we will end up with many possible candidate density functions so the last step involves a hypothesis test to select a candidate that is indeed close to the true distribution in TV distance.

Algorithm 1 FULL ALGORITHM

```
Input: $\epsilon$-corrupted sample $X_1, \dots, X_n$ from mixture of Gaussians $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$
LEARN PARAMETERS TO $\epsilon^{\Omega(1)}$ ACCURACY for each set of candidate components $\overline{G_1}, \dots \overline{G_k}$ do
    Assign samples to components according to maximum likelihood to form sets of samples
    $\{\overline{S_1}, \dots, \overline{S_k}\}$ for all partitions of $[k]$ into sets $R_1, \dots, R_l$ do
        LEARN MIXTURE TO $\tilde{O}(\epsilon)$ ACCURACY on samples $\mathcal{R}_j = \cup_{i \in R_j} \overline{S_i}$ with initial estimates
        $\{\overline{G_i}\}_{i \in R_j}$
    end
    Compute density estimate by combining over all of $R_1, \dots, R_l$ and guessing weights of each
    submixture
end
Hypothesis test over all candidate density estimates to output an estimate $f$ that is $\tilde{O}(\epsilon)$-close to
$\mathcal{M}$
```

The algorithm LEARN MIXTURE TO $\tilde{O}(\epsilon)$ ACCURACY requires that the components of the mixture are not too far from each other in TV distance (so that there exists a transformation that puts the mixture in regular form) and also requires initial estimates $\overline{G_1}, \dots, \overline{G_k}$ for the component Gaussians that are $\epsilon^{\Omega(1)}$ close to the true components in TV distance so that we can apply Theorem 69. We will show that the clustering step ensures the first property. The initial estimates are simply obtained from the output of the first step of LEARN PARAMETERS TO $\epsilon^{\Omega(1)}$ ACCURACY.

Algorithm 2 LEARN MIXTURE TO $\tilde{O}(\epsilon)$ ACCURACY

```
Input: $\epsilon$-corrupted sample $X_1, \dots, X_n$ from mixture of Gaussians $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$
such that $d_{\mathsf{TV}}(G_i, G_j) \leq 1 - \epsilon^{O(1)}$ for all $i \neq j$
Input: Initial estimates $\overline{G_1}, \dots, \overline{G_k}$ such that for all $i \in [k]$,

    $$d_{\mathsf{TV}}(\overline{G_i}, G_i) \leq \epsilon^{\Omega(1)} .$$

Let $\widetilde{\mu}_1, \widetilde{\Sigma}_1$ be the mean and covariance of $\overline{G_1}$.
Apply the transformation $X_i \to \widetilde{\Sigma}_1^{-1/2}(X_i - \widetilde{\mu}_1)$ to the datapoints
Use Theorem 69 on the transformed data to compute a density estimate $f$
Output density estimate $f(\widetilde{\Sigma}_1^{-1/2}(x - \widetilde{\mu}_1)) \cdot \det(\Sigma_1)^{-1/2}$
```

The main theorem of this paper, which we prove in this section, is stated below.

Theorem 72 *Let $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ be a χ-balanced mixture of Gaussians (recall Definition 70). Furthermore, assume that all of the mixing weights are at least A^{-1} for some constant A. Assume that ϵ is sufficiently small compared to k, A and $\chi \leq \mathrm{poly}(\log 1/\epsilon)$. Let $n = \mathrm{poly}_{k,A}(d/\epsilon)$ for some sufficiently large polynomial and let $X_1, \dots, X_n$ be an ϵ-corrupted sample from $\mathcal{M}$. Then* FULL ALGORITHM *runs in $\mathrm{poly}_{k,A}(d/\epsilon)$ time and with 0.9 probability, outputs a degree $O_{k,A}(1)$*

MPG distribution f such that

$$\|f(x) - \mathcal{M}(x)\|_1 \leq (2 + \log 1/\epsilon + \chi)^{O_{k,A}(1)}\epsilon\,.$$

I.1. Estimate Components to $\epsilon^{\Omega(1)}$ Accuracy

First, we will estimate the components of the mixture to $\epsilon^{\Omega(1)}$ accuracy. This can be done with a few simple modifications to the techniques in Liu and Moitra (2021).

Theorem 73 *Let $k, A > 0$ be constants. There is a sufficiently large function G and a sufficiently small function g depending only on k, A such that given an ϵ-corrupted sample $X_1, \dots, X_n$ from a mixture of Gaussians $\mathcal{M} = w_1 G_1 + \dots + w_k G_k \in \mathbb{R}^d$ where $\epsilon < g(k, A)$, the w_i are all at least A^{-1}, and $n \geq (d/\epsilon)^{G(k,A)}$, there is an algorithm that runs in time $\text{poly}(n)$ and with 0.999 probability, outputs a set of $(1/\epsilon)^{O_{k,A}(1)}$ candidate mixtures such that for at least one of these candidates, $\{\widetilde{w_1}, \overline{G_1}, \dots, \widetilde{w_k}, \overline{G_k}\}$, we have*

$$|w_i - \widetilde{w_i}| + d_{\mathsf{TV}}(G_i, \overline{G_i}) \leq \epsilon^{g(k,A)}$$

for all $i \in [k]$.

I.1.1. ACHIEVING CONSTANT ACCURACY

First, we will show that we can estimate the components to constant accuracy. Recall the following result from Liu and Moitra (2021).

Lemma 74 *[Lemma 7.5 in Liu and Moitra (2021)] Let $k, A, b > 0$ be constants and θ be a desired accuracy. There is a sufficiently large function G and a sufficiently small function g depending only on k, A, b, θ such that given an ϵ-corrupted sample $X_1, \dots, X_n$ from a mixture of Gaussians $\mathcal{M} = w_1 G_1 + \dots + w_k G_k \in \mathbb{R}^d$ where*

- *The w_i are all at least A^{-1} for some constant A*
- *$d_{\mathsf{TV}}(G_i, G_j) \geq b$*

and

- *$\epsilon < g(k, A, b, \theta)$*
- *$n \geq (d/\epsilon)^{G(k,A,b,\theta)}$*

then there is an algorithm that runs in time $\text{poly}(n)$ and with 0.999 probability outputs a set of $(1/\theta)^{G(k,A,b,\theta)}$ candidate mixtures at least one of which satisfies

$$\max\left(d_{\mathsf{TV}}(\overline{G_1}, G_1), \dots, d_{\mathsf{TV}}(\overline{G_k}, G_k)\right) \leq \theta$$
$$\max\left(|\widetilde{w_1} - w_1|, \dots, |\widetilde{w_k} - w_k|\right) \leq \theta$$

Remark 75 *Lemma 7.5 in Liu and Moitra (2021) is stated with an assumption that the w_i have bounded fractionality with denominator at most A. The modification given in Section 6.3 in Liu and Moitra (2021) (namely Theorem 6.12) immediately allows us to remove the bounded fractionality part of the assumption.*

Fix k, A. Note that we would like to eliminate the assumption that $d_{\mathsf{TV}}(G_i, G_j) \geq b$ in the above result. In fact, for achieving constant accuracy, this is not difficult to do because we can simply find some scale for which we can lump together components whose TV distance is too small and otherwise the components will be sufficiently separated.

Corollary 76 *Let k, A be constants. Let θ be some desired accuracy. There is a sufficiently large function F and a sufficiently small function f depending only on k, A, θ (with $F(k, A, \theta), f(k, A, \theta) > 0$) such that the following holds. Given an ϵ-corrupted sample $X_1, \dots, X_n$ from a mixture of Gaussians $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k \in \mathbb{R}^d$ where*

- *The w_i are all at least A^{-1} for some constant A*
- $\epsilon < f(k, A, \theta)$
- $n \geq (d/\epsilon)^{F(k,A,\theta)}$

then there is an algorithm that runs in time $\mathrm{poly}(n)$ and with 0.999 probability outputs a set of $(1/\theta)^{F(k,A,\theta)}$ candidate mixtures at least one of which satisfies

$$\max\left(d_{TV}(\overline{G_1}, G_1), \dots, d_{TV}(\overline{G_k}, G_k)\right) \leq \theta$$
$$\max\left(|\widetilde{w_1} - w_1|, \dots, |\widetilde{w_k} - w_k|\right) \leq \theta$$

Proof Let g and G be the functions in Lemma 74. Fix k, A, θ. We define $h(b) = 0.1\theta g(k, A, b, 0.5\theta)/k$. Consider the sequence

$$\theta \to h(\theta) \to \cdots \to h^{(k^2)}(\theta)\,.$$

There must be some $j < k^2$ such that no pair of true components has TV distance between $h^{(j)}(\theta)$ and $h^{j+1}(\theta)$. Now consider the graph $\mathcal{G}$ on $[k]$ where two indices i_1, i_2 are connected if and only if

$$d_{\mathsf{TV}}(G_{i_1}, G_{i_2}) \leq h^{j+1}(\theta)\,.$$

Now consider a modified mixture $\mathcal{M}'$ where for each connected component of vertices in $\mathcal{G}$, say $S \subset [k]$, we replace all of the Gaussians G_i with $i \in S$ with copies of one fixed representative from this component. We then combine all of these copies by adding the mixing weights. The resulting mixture $\mathcal{M}'$ satisfies the following properties

- There are $k' \leq k$ components
- All mixing weights are at least A^{-1}
- All pairs of components are separated by at least $h^{(j)}(\theta)$ in TV distance
- $$d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leq 0.1\theta g(k, A, h^{(j)}(\theta), 0.5\theta)$$

In particular, we can treat our samples as an ϵ'-corrupted sample from $\mathcal{M}'$ with

$$\epsilon' = \epsilon + 0.1\theta g(k, A, h^{(j)}(\theta), 0.5\theta)\,.$$

Now as long as f, F are chosen appropriately, we can apply Lemma 74 to learn the components of the mixture $\mathcal{M}'$ to accuracy 0.5θ. Finally, we can simply guess the mixing weights and duplications in our list of candidates to ensure that one of our candidate mixtures is component-wise within θ of the true mixture. ■

I.1.2. ACHIEVING $\epsilon^{\Omega(1)}$-ACCURACY

Similar to Liu and Moitra (2021), once we obtain constant accuracy estimates for the components, we then try to refine these estimates. To do this, we first cluster the datapoints into submixtures by assigning each datapoint to the submixture that assigns it the highest probability. We would like the following two properties

- The clustering is $1 - \epsilon^c$-accurate
- Components within a submixture have TV distance at most $1 - \epsilon^{c'}$

where c' is sufficiently small compared to c. Once we have clustered the datapoints into such submixtures, we can learn the components of each submixture to $\epsilon^{\Omega(1)}$ accuracy, again following the same outline as the algorithm in Liu and Moitra (2021).

First we show that the clustering can be done accurately. We need the following basic results from Liu and Moitra (2021).

Lemma 77 (Lemma 7.2 from Liu and Moitra (2021)) *Let A, B, C be Gaussian distributions. Assume that $d_{\mathsf{TV}}(A, B) \leq 0.9$. There is a universal constant $c > 0$ such that if $d_{\mathsf{TV}}(A, C) \geq 1 - \epsilon$ and $\epsilon < c$ then*

$$d_{\mathsf{TV}}(B, C) \geq 1 - \epsilon^c \,.$$

Lemma 78 (Lemma 7.4 from Liu and Moitra (2021)) *Let A and B be two Gaussians with $d_{\mathsf{TV}}(A, B) \leq 0.9$. There is a universal constant $c > 0$ such that if $D \in \{A, B\}$ and $\epsilon < c$ then*

$$P_{x \sim D}\left[\epsilon \leq \frac{A(x)}{B(x)} \leq \frac{1}{\epsilon}\right] \geq 1 - \epsilon^c \,.$$

Similar to Liu and Moitra (2021), we will also need VC-dimension bounds on the hypothesis class formed by comparing two MPG functions. For the clustering step, we only need to deal with actual mixtures of Gaussians, but we will need the VC-dimension bound for MPG functions later on when we do hypothesis testing so we state the full result here. First we need a definition.

Definition 79 *Let $\mathcal{F}$ be a family of functions on some domain $\mathcal{X}$. Let $\mathcal{H}_{\mathcal{F},a}$ be the set of functions of the form $f_{\mathcal{M}_1,\mathcal{M}_2,\dots,\mathcal{M}_a}$ where $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_a \in \mathcal{F}$ and*

$$f_{\mathcal{M}_1,\mathcal{M}_2,\dots,\mathcal{M}_a}(x) = \begin{cases} 1 \text{ if } \mathcal{M}_1(x) \geq \mathcal{M}_2(x), \dots, \mathcal{M}_a(x) \\ 0 \text{ otherwise} \end{cases}$$

The VC dimension bound below is a direct consequence of the work in Anthony and Bartlett (2009).

Lemma 80 (Theorem 8.14 in Anthony and Bartlett (2009)) *Let $\mathcal{F}_{k,m}$ be the family of functions in $\mathbb{R}^d$ that are a degree m MPG function with at most k components. Then the VC dimension of $\mathcal{H}_{\mathcal{F}_{k,m},a}$ is* $\mathrm{poly}(d, a, m, k)$.

It is a standard result in learning theory that for a hypothesis class with bounded VC dimension, taking a polynomial number of samples suffices to get a good approximation for all hypotheses in the class.

Lemma 81 (Vapnik and Chervonenkis (2015)) *Let $\mathcal{H}$ be a hypothesis class of functions from some domain $\mathcal{X}$ to $\{0,1\}$ with VC dimension V. Let $\mathcal{D}$ be a distribution on $\mathcal{X}$. Let $\epsilon, \delta > 0$ be parameters. Let S be a set of $n = \text{poly}(V, 1/\epsilon, \log 1/\delta)$ i.i.d samples from $\mathcal{D}$. Then with $1-\delta$ probability, for all $f \in \mathcal{H}$*

$$|\mathbb{E}_{x\sim S}[f(x)] - \mathbb{E}_{x\sim\mathcal{D}}[f(x)]| \leq \epsilon\,.$$

Now we prove that given constant-accuracy component estimates, we can find an accurate clustering.

Lemma 82 *Let $\mathcal{M} = w_1G_1 + \cdots + w_kG_k \in \mathbb{R}^d$ be a mixture of Gaussians where the w_i are all at least A^{-1} for some constant A. There exists a sufficiently small function $g(k,A) > 0$ depending only on k, A such that the following holds. Let $X_1, \dots, X_n$ be an ϵ-corrupted sample from the mixture $\mathcal{M}$ where $\epsilon < g(k,A)$ and $n = \text{poly}_{k,A}(d/\epsilon)$ for some sufficiently large polynomial. Let $S_1, \dots, S_k \subset \{X_1, \dots, X_n\}$ denote the sets of samples from each of the components $G_1, \dots, G_k$ respectively. Let $R_1, \dots, R_l$ be a partition of $[k]$ such that for $i_1 \in R_{j_1}, i_2 \in R_{j_2}$ with $j_1 \neq j_2$,*

$$d_{\mathsf{TV}}(G_{i_1}, G_{i_2}) \geq 1 - \epsilon'$$

where $\epsilon' \leq g(k,A)$. Let $\overline{G_1}, \dots, \overline{G_k}$ be any Gaussians such that $d_{\mathsf{TV}}(G_i, \overline{G_i}) \leq g(k,A)$ for all i. Let $\overline{S_1}, \dots, \overline{S_k} \subset \{X_1, \dots, X_n\}$ be the subsets of samples obtained by assigning each sample to the component $\overline{G_i}$ that gives it the maximum likelihood. Then there is a universal constant $\eta > 0$ such that with probability at least 0.999,

$$\left|\left(\cup_{i\in R_j} S_i\right) \cap \left(\cup_{i \in R_j} \overline{S_i}\right)\right| \geq (1 - O_{k,A}(1)\epsilon - \epsilon'^{\eta}) \max\left(\left|\left(\cup_{i\in R_j} S_i\right)\right|, \left|\left(\cup_{i\in R_j}\overline{S_i}\right)\right|\right)$$

for all $j \in [l]$.

Proof We will upper bound the expected number of uncorrupted points that are mis-classified for each $j \in [l]$ and then use the VC dimension bound in Lemma 80 and the uniform convergence guarantee in Lemma 81 to argue that for any initial estimates $\overline{G_1}, \dots, \overline{G_k}$, the fraction of points that we mis-classify is small. The expected number of uncorrupted points can be upper bounded by

$$\sum_{j_1 \neq j_2} \sum_{\substack{i_1 \in R_{j_1} \\ i_2 \in R_{j_2}}} \int 1_{\overline{G_{i_1}}(x) > \overline{G_{i_2}}(x)} dG_{i_2}(x)\,.$$

Clearly we can ensure $d_{\mathsf{TV}}(G_i, \overline{G_i}) \leq 1/2$. Thus, by Lemma 77 and the assumption about $R_1, \dots, R_l$, $d_{\mathsf{TV}}(\overline{G_{i_1}}, G_{i_2}) \geq 1 - \epsilon'^{\Omega(1)}$ for all G_{i_2} where i_2 is not in the same piece of the partition as i_1. Let c be such that

$$d_{\mathsf{TV}}(\overline{G_{i_1}}, G_{i_2}) \geq 1 - \epsilon'^{c}\,.$$

By Lemma 78,

$$\Pr_{x \in G_{i_2}}\left[\epsilon'^{c/2} \leq \frac{\overline{G_{i_2}}(x)}{G_{i_2}(x)} \leq \epsilon'^{c/2}\right] \geq 1 - \epsilon'^{\Omega(c)}$$

and combining the above two inequalities, we deduce

$$\int 1_{\overline{G_{i_1}}(x) > \overline{G_{i_2}}(x)} dG_{i_2}(x) \leq \epsilon'^{\Omega(1)}\,.$$

Since we are only summing over $O_k(1)$ pairs of components, as long as ϵ' is sufficiently small compared to k, A, the expected fraction of misclassified uncorrupted points is $\epsilon'^{\Omega(1)}$.

It remains to note that the clustering depends only on the comparisons between the values of the pdfs of the Gaussians $\overline{G_1}, \dots, \overline{G_k}$ at each of the samples $X_1, \dots, X_n$. Since $n = \text{poly}_{k,A}(d/\epsilon)$ for some sufficiently large polynomial, applying Lemma 80 and Lemma 81 completes the proof (note that the fraction of corrupted points is at most ϵ overall and the mixing weights are lower bounded so the fraction of corrupted points in each cluster is at most $O_{k,A}(1)\epsilon$). ■

Finally, it remains to note that once we have clustered the points, we can learn the parameters. For this, we rely on the following definition and result from Liu and Moitra (2021).

Definition 83 (Definition 5.1 in Liu and Moitra (2021)) *We say a mixture of Gaussians $w_1 G_1 + \dots + w_k G_k$ is δ-tight if*

1. *Let $\mathcal{G}$ be the graph on $[k]$ obtained by connecting two nodes i, j if $d_{\mathsf{TV}}(G_i, G_j) \leq 1 - \delta$. Then $\mathcal{G}$ is connected*
2. *$d_{\mathsf{TV}}(G_i, G_j) \geq \delta$ for all $i \neq j$*
3. *$w_{\min} \geq \delta$*

Theorem 84 (Theorem 5.2 in Liu and Moitra (2021)) *There is a function $f(k) > 0$ depending only on k such that given an ϵ-corrupted sample from a δ-tight mixture of Gaussians*

$$\mathcal{M} = w_1 N(\mu_1, \Sigma_1) + \dots + w_k N(\mu_k, \Sigma_k)$$

where $\delta \geq \epsilon^{f(k)}$, there is a polynomial time algorithm that outputs a set of $(1/\epsilon)^{O_k(1)}$ candidate mixtures $\{\widetilde{w_1} N(\widetilde{\mu_1}, \widetilde{\Sigma_1}) + \dots + \widetilde{w_k} N(\widetilde{\mu_k}, \widetilde{\Sigma_k}\}$ and with high probability, at least one of them satisfies that for all i:

$$|w_i - \widetilde{w_i}| + d_{\mathsf{TV}}(N(\mu_i, \Sigma_i), N(\widetilde{\mu_i}, \widetilde{\Sigma_i})) \leq \epsilon^{\Omega_k(1)}\,.$$

With one additional argument, we can remove the assumption that the components are δ-separated in TV distance from the above.

Corollary 85 *There is a function $g(k) > 0$ depending only on k such that given an ϵ-corrupted sample from a mixture of Gaussians*

$$\mathcal{M} = w_1 N(\mu_1, \Sigma_1) + \dots + w_k N(\mu_k, \Sigma_k)$$

where $\mathcal{M}$ satisfies conditions 1 and 3 of Definition 83 for some $\delta \geq \epsilon^{g(k)}$, there is a polynomial time algorithm that outputs a set of $(1/\epsilon)^{O_k(1)}$ candidate mixtures $\{\widetilde{w_1} N(\widetilde{\mu_1}, \widetilde{\Sigma_1}) + \dots + \widetilde{w_k} N(\widetilde{\mu_k}, \widetilde{\Sigma_k}\}$ and with high probability, at least one of them satisfies that for all i:

$$|w_i - \widetilde{w_i}| + d_{\mathsf{TV}}(N(\mu_i, \Sigma_i), N(\widetilde{\mu_i}, \widetilde{\Sigma_i})) \leq \epsilon^{\Omega_k(1)}\,.$$

Proof Let $f(k)$ be the function in Theorem 84. Consider the sequence

$$\psi_0 = \epsilon, \psi_1 = \epsilon^{(0.1f(k))}, \dots, \psi_{k^2} = \epsilon^{(0.1f(k))^{k^2}} .$$

There must be an index $j < k^2$ such that there is no pair of components G_{i_1} and G_{i_2} with TV distance between ψ_j and ψ_{j+1}. Let $\mathcal{G}$ be the graph on $[k]$ where two indices i_1 and i_2 are connected if and only if

$$d_{\mathsf{TV}}(G_{i_1}, G_{i_2}) \leq \psi_j .$$

Now consider a modified mixture $\mathcal{M}'$ where for each connected component of vertices in $\mathcal{G}$, say $S \subset [k]$, we replace all of the Gaussians G_i with $i \in S$ with copies of one fixed representative from this component. We then combine the copies by adding the mixing weights. The resulting mixture $\mathcal{M}'$ satisfies the following properties

- $\mathcal{M}'$ has $k' \leq k$ components $G'_1, \dots, G'_{k'}$
- The graph on $[k']$ obtained by connecting two nodes i, j if $d_{\mathsf{TV}}(G'_i, G'_j) \leq 1 - \delta + k\psi_j$ is connected
- All pairs of components are separated by at least ψ_{j+1} in TV distance
- All mixing weights are at least δ
- $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leq k\psi_j$

Thus, we can treat our sample as an ϵ'-corrupted sample from $\mathcal{M}'$ with

$$\epsilon' = \epsilon + k\psi_j .$$

Note that $\psi_{j+1} \geq \epsilon'^{f(k)}$ and thus by choosing $g(k)$ appropriately, we can ensure that the mixture $\mathcal{M}'$ is $\epsilon'^{f(k)}$-tight. Thus, we can apply Theorem 84 to learn the components of the mixture $\mathcal{M}'$ to within $\epsilon'^{\Omega_k(1)} = \epsilon^{\Omega_k(1)}$. We can then guess the mixing weights and duplications of components to output a list of candidate mixtures at least one of which is component-wise within $\epsilon^{\Omega_k(1)}$ of the true mixture. ■

To put everything together, we will need the following simple claim which also appears in Liu and Moitra (2021).

Claim 37 (Claim 7.6 in Liu and Moitra (2021)) *Let $\mathcal{M} = w_1 G_1 + \dots + w_k G_k$ be a mixture of Gaussians. For any constant $c > 0$ and parameter ϵ, there exists a function $f(c, k)$ such that there exists a partition (possibly trivial) of $[k]$ into sets $R_1, \dots, R_l$ such that*

- *If we draw edges between all i, j such that $d_{\mathsf{TV}}(G_i, G_j) \leq 1 - \epsilon^{c\kappa}$ then each piece of the partition is connected*
- *For any i, j in different pieces of the partition $d_{\mathsf{TV}}(G_i, G_j) \geq 1 - \epsilon^{\kappa}$*

and $f(c, k) < \kappa < 1$.

Now we can prove Theorem 73.

Proof [Proof of Theorem 73] This follows from combining Corollary 76, Claim 37, Lemma 82 and finally applying Corollary 85. Note we can choose the constant c in Claim 37 sufficiently small so that when combined with Lemma 82, the resulting accuracy that we get on each submixture is high enough that we can then apply Corollary 85 (we can treat the subsample corresponding to each submixture as a $O(\epsilon'^{\eta})$-corrupted sample from that submixture). We apply Lemma 82 with $\epsilon' = \epsilon^{\kappa}$ where the κ is obtained from Claim 37. ■

I.2. Estimate Mixture to $\tilde{O}(\epsilon)$ accuracy

Note that the results in the previous section do not require the Gaussians to be reasonably well-conditioned. However for the next step, of going from $\epsilon^{\Omega(1)}$ accuracy to $\tilde{O}(\epsilon)$ accuracy, we will require the assumption that the Gaussians are reasonably well-conditioned. We will restrict to the case where the Gaussians are χ-balanced where χ should be thought of as a constant (or $\mathrm{poly}(\log 1/\epsilon)$).

The main theorem that we will prove in this section is as follows.

Theorem 86 *Let $\mathcal{M} = w_1 G_1 + \cdots + w_k G_k$ be a χ-balanced mixture of Gaussians. Furthermore, assume that all of the mixing weights are at least A^{-1} for some constant A. Assume that ϵ is sufficiently small compared to k, A and $\chi \leq \mathrm{poly}(\log 1/\epsilon)$. Let $n = \mathrm{poly}_{k,A}(d/\epsilon)$ for some sufficiently large polynomial and let $X_1, \dots, X_n$ be an ϵ-corrupted sample from $\mathcal{M}$. Then with 0.99 probability, the list of candidate density estimates computed by* FULL ALGORITHM *is a list of size $(1/\epsilon)^{O_{k,A}(1)}$ and contains a degree $O_{k,A}(1)$ MPG distribution f such that*

$$\|f(x) - \mathcal{M}(x)\|_1 \leq (2 + \log 1/\epsilon + \chi)^{O_{k,A}(1)} \epsilon\,.$$

Proof Let $g(k, A)$ be the function in Theorem 73. Using Theorem 73, with 0.999 probability, among the list of candidate components computed in LEARN PARAMETERS TO $\epsilon^{\Omega(1)}$ ACCURACY, there is some set $\{\overline{G_1}, \dots, \overline{G_k}\}$ such that for all $i \in [k]$

$$d_{\mathsf{TV}}(\overline{G_i}, G_i) \leq \epsilon^{g(k,A)}\,.$$

Now consider the graph $\mathcal{G}$ on $[k]$ where two nodes i and j are connected if and only if $d_{\mathsf{TV}}(G_i, G_j) \leq 1 - \epsilon^{1/\eta}$ where η is the universal constant in Lemma 82. Let $R_1, \dots, R_l \subset [k]$ be the connected components in this graph. For each $j \in [l]$ define the submixture

$$\mathcal{M}_j = \frac{\sum_{i \in R_j} w_i G_i}{\sum_{i \in R_j} w_i}\,.$$

Now by Lemma 82, with 0.999 probability, if we partition $[k]$ according to $R_1, \dots, R_l$ and then assign samples by maximum likelihood using the estimates $\{\overline{G_1}, \dots, \overline{G_k}\}$, the resulting the subsets of samples $\mathcal{R}_1, \dots, \mathcal{R}_l$ are equivalent to $O_{k,A}(1)\epsilon$-corrupted samples from each of the submixtures $\mathcal{M}_1, \dots, \mathcal{M}_l$ (note that we are setting $\epsilon' = \epsilon^{1/\eta}$ in Lemma 82). It now suffices to estimate the density function of each submixture to $\widetilde{O}(\epsilon)$ accuracy. We will argue that given this clustering, LEARN MIXTURE TO $\widetilde{O}(\epsilon)$ ACCURACY indeed learns each of the submixtures to $\widetilde{O}(\epsilon)$ accuracy.

Let $\mu_1, \Sigma_1, \dots \mu_k, \Sigma_k$ be the means and covariances of the true components $G_1, \dots, G_k$ and let $\widetilde{\mu}_1, \widetilde{\Sigma}_1, \dots, \widetilde{\mu}_k, \widetilde{\Sigma}_k$ be the means and covariances of our initial estimates. Without loss of generality assume $\mathcal{R}_1 = \{1, 2, \dots, t\}$. Let L denote the linear transformation $x \to \widetilde{\Sigma}_1^{-1/2}(x - \widetilde{\mu}_1)$. Since $d_{\mathsf{TV}}(\overline{G}_i, G_i) \leq \epsilon^{g(k,A)}$, the Gaussians $\overline{G}_i$ must all be 2χ-balanced. Now by Claim 36, we conclude

$$\|L(\mu_1)\| + \|L(\Sigma_1) - I\|_2 \leq \epsilon^{\Omega_{k,A}(1)} .$$

Also since $\mathcal{R}_1 = \{1, 2, \dots, t\}$ is a connected a component in $\mathcal{G}$, we can apply Claim 35 and sum over all paths in the connected component to deduce that

$$\|L(\mu_i)\| , \|L(\Sigma_i) - I\|_2 \leq \mathrm{poly}(\chi, \log 1/\epsilon)$$

for all $i \leq t$. Thus, since ϵ is sufficiently small in terms of k, A, we can ensure that the transformed mixture $L(\mathcal{M}_1)$ is in (α, β, γ)-regular form for

- $\alpha \leq \mathrm{poly}(\log 1/\epsilon)$
- $\beta \leq \mathrm{poly}(\log 1/\epsilon)$
- γ sufficiently small in terms of k, A

Now the above implies that the application of Theorem 69 in algorithm LEARN MIXTURE TO $\widetilde{O}(\epsilon)$ ACCURACY is valid (with $c = \Omega_{k,A}(1)$) and guarantees that with high probability, after applying the inverse linear transformation, we obtain a function f_1 such that

$$\|f_1(x) - \mathcal{M}_1(x)\|_1 \leq (2 + \log 1/\epsilon + \chi)^{O_{k,A}(1)}\epsilon .$$

Similarly, we compute functions $f_2, \dots, f_l$ that approximate the density functions of $\mathcal{M}_2, \dots, \mathcal{M}_l$. Finally note that

$$\mathcal{M} = \sum_{j=1}^{l} \left(\sum_{i \in R_j} w_i \right) \mathcal{M}_j .$$

We can simply guess the weights $\left(\sum_{i \in R_j} w_i\right)$ for $j \in [l]$ of the submixtures using an ϵ-grid. If our guesses, say $W_1, \dots, W_l$ are all within ϵ of the true values then the function $f = \sum_{j=1}^{l} W_j f_j(x)$ satisfies the desired conditions because

$$\left\| \sum_{j=1}^{l} W_j f_j(x) - \mathcal{M}(x) \right\|_1 \leq \left\| \sum_{j=1}^{l} W_j f_j(x) - \sum_{j=1}^{l} W_j \mathcal{M}_j \right\|_1 + \left\| \sum_{j=1}^{l} \left(W_j - \sum_{i \in R_j} w_i \right) \mathcal{M}_j \right\|_1$$
$$\leq (2 + \log 1/\epsilon + \chi)^{O_{k,A}(1)}\epsilon .$$

Overall, the number of candidates that FULL ALGORITHM outputs is $(1/\epsilon)^{O_{k,A}(1)}$ (enumerating over all initial estimates for the components, possible clusterings, and possible guesses for the weights) and with 0.99 probability at least one of them satisfies

$$\|f(x) - \mathcal{M}(x)\|_1 \leq (2 + \log 1/\epsilon + \chi)^{O_{k,A}(1)}\epsilon ,$$

completing the proof. ■

I.3. Hypothesis Testing

Theorem 86 guarantees that we can learn a list of candidate density functions at least one of which is close to the density function of the true mixture. The last step is to hypothesis test to select an element of the list which is guaranteed to be close to the true density function. For this we rely on the following lemma from Liu and Moitra (2021).

Lemma 87 (Restated from Liu and Moitra (2021)) *Let $\mathcal{F}$ be a family of distributions on some domain $\mathcal{X}$ with explicitly computable density functions that can be efficiently sampled from. Let V be the VC dimension of $\mathcal{H}_{\mathcal{F},2}$ (recall Definition 79). Let $\mathcal{D}$ be an unknown distribution in $\mathcal{F}$. Let m be a parameter. Let $X_1, \dots, X_n$ be an ϵ-corrupted sample from $\mathcal{D}$ with $n \geq \mathrm{poly}(m, \epsilon, V)$ for some sufficiently large polynomial. Let $H_1, \dots, H_m$ be distributions in $\mathcal{F}$ given to us by an adversary with the promise that*

$$\min(d_{\mathsf{TV}}(\mathcal{D}, H_i)) \leq \epsilon \,.$$

Then there exists an algorithm that runs in time $\mathrm{poly}(n, \epsilon)$ and outputs an i with $1 \leq i \leq m$ such that with 0.999 probability

$$d_{TV}(\mathcal{D}, H_i) \leq O(\epsilon) \,.$$

We can now complete the proof of the main theorem.

Proof [Proof of Theorem 72] Combining Theorem 86 with Lemma 87 and Lemma 80, we immediately get the desired result. Note that the density function of a constant degree MPG distribution is explicitly computable. To see why it can be efficiently sampled from, note that all of the polynomials are always positive and the integral of $\int_{\mathbb{R}^d} P(x)G(x)dx$ for a polynomial P and Gaussian G can be computed explicitly using integration by parts. ∎

Appendix J. Omitted Proofs from Section F

Proof [Proof of Claim 28] First, we obtain high probability bounds on the sum of the largest ϵ-fraction of the samples.

For any α, let S_α be the set of $x \in S$ with $|x| \geq \alpha \log^{1/c}(1/\epsilon)$. Let $C = (10/c)^{10/c}$. For $C \leq \alpha \leq n^{0.01}$, using tail bounds on the binomial distribution, we get

$$\Pr\left[|S_\alpha| > \frac{\epsilon n}{\alpha^{10}}\right] \leq \binom{n}{\epsilon n/\alpha^{10}} \epsilon^{\alpha^c \cdot \frac{n\epsilon}{\alpha^{10}}} \leq \left(\frac{3\alpha^{10}}{\epsilon} \cdot \epsilon^{\alpha^c}\right)^{\frac{n\epsilon}{\alpha^{10}}} \leq \frac{e^{-(10d/\epsilon)^2}}{\alpha^{10}} \,.$$

Also, using a simple union bound, with probability at least $1 - e^{-(10d/\epsilon)^2}$, we have $|x| \leq n^{0.01}$ for all $x \in S$.

Consider a set of α forming a geometric series with ratio 2, say $\{C, 2C, \dots, \}$. Combining everything we've shown so far using a union bound, with probability at least $1 - e^{-(9d/\epsilon)^2}$, for all $\alpha = \{C, 2C, \dots, \}$, we have

$$|S_\alpha| \leq \frac{\epsilon n}{\alpha^{10}} \,.$$

This implies that with probability $1 - e^{-(9d/\epsilon)^2}$, for any set $T \subset S$ of size at most ϵn,

$$\sum_{x \in T} |x| \leq \epsilon C \log^{1/c}(1/\epsilon) n + \sum_{i=1}^{\infty} \frac{2^i C}{2^{10(i-1)}} \log^{1/c}(1/\epsilon)\epsilon n \leq 10\epsilon \log^{1/c}(1/\epsilon) C n \,. \tag{20}$$

Let $\mathcal{D}'$ be the distribution $\mathcal{D}$ restricted to the interval $\left[-\log^{1/c}(1/\epsilon), \log^{1/c}(1/\epsilon)\right]$. The assumption about the tail decay of $\mathcal{D}$ implies that

$$|\mu_{\mathcal{D}'} - \mu_{\mathcal{D}}| \leq 10\epsilon \log^{1/c}(1/\epsilon) C \,. \tag{21}$$

Let λ be the probability mass that $\mathcal{D}$ has in the interval $\left[-\log^{1/c}(1/\epsilon), \log^{1/c}(1/\epsilon)\right]$. Sampling from $\mathcal{D}$ is equivalent to sampling using the following procedure.

- First draw a Bernoulli random variable σ that is 1 with probability λ
- If σ is 1 then draw a sample from $\mathcal{D}'$ and otherwise sample from the distribution $\mathcal{D}$ restricted to outside the interval $\left[-\log^{1/c}(1/\epsilon), \log^{1/c}(1/\epsilon)\right]$

Note that $\lambda \geq 1 - \epsilon$. Assuming that the original set of samples S is drawn in this manner, with probability at least $1 - e^{-(9d/\epsilon)^2}$, there are at least $(1-2\epsilon)n$ elements of S that are drawn from $\mathcal{D}'$. By a Chernoff bound, their mean is within $10\epsilon \log^{1/c}(1/\epsilon)C$ of $\mu_{\mathcal{D}'}$ with at least $1 - e^{-(9d/\epsilon)^2}$ probability. Finally, using equations (20), (21) and the triangle inequality, we deduce that with probability at least $1 - e^{-(8d/\epsilon)^2}$, for all subset $S' \subset S$ of size at least $(1-\epsilon)n$,

$$\left| \mu_{\mathcal{D}} - \frac{1}{|S'|}\sum_{x \in S'} x \right| \leq \epsilon \log^{1/c}(1/\epsilon) \left(\frac{10^2}{c}\right)^{10/c} .$$

■

Proof [Proof of Claim 29] Let κ be the dimensionality of $v(H_m(X, z))$. Note $\kappa = d^{O_m(1)}$. Let E be an $\epsilon/(10\kappa)$-net of the unit sphere in κ dimensions. We can ensure that

$$|E| \leq \left(\frac{10^2 \kappa}{\epsilon}\right)^{\kappa} .$$

For a fixed vector $v \in E$, we will compute the probability that either

$$\left| v \cdot \left(\mu_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} x \right) \right| \geq \frac{\delta}{2}$$

or

$$\left| v^T \left(\Sigma_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} (x - \mu_{\mathcal{D}})(x - \mu_{\mathcal{D}})^T \right) v \right| \geq \frac{\delta^2}{2\epsilon}$$

for some subset S' with $|S'| \geq (1-\epsilon)|S|$ and then we will union bound the failure probability over all vectors $v \in E$.

By Lemma 61, the distribution $v \cdot \mathcal{D}$, scaled by a factor of

$$\frac{1}{(2+\alpha+\beta)^{O_m(1)}} \,,$$

satisfies the conditions of Claim 28 with $c = \Omega_m(1)$. Thus, by choosing n sufficiently large, we can ensure that with $1 - e^{(8\kappa/\epsilon)^2}$ probability, we have

$$\left| v \cdot \left(\mu_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} x \right) \right| \leq \frac{\delta}{2}$$

for all subsets S' with $|S'| \geq (1-\epsilon)n$. Next,

$$\begin{aligned} &\left| v^T \left(\Sigma_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} (x - \mu_{\mathcal{D}})(x - \mu_{\mathcal{D}})^T \right) v \right| \\ &= \left| v^T (\Sigma_{\mathcal{D}} + \mu_{\mathcal{D}} \mu_{\mathcal{D}}^T) v - \frac{1}{S'} \sum_{x \in S'} (v \cdot x)^2 + \frac{2}{S'} v^T \left(\sum_{x \in S'} (x - \mu_{\mathcal{D}}) \right) \mu_{\mathcal{D}}^T v \right| \\ &\leq \left| v^T (\Sigma_{\mathcal{D}} + \mu_{\mathcal{D}} \mu_{\mathcal{D}}^T) v - \frac{1}{S'} \sum_{x \in S'} (v \cdot x)^2 \right| + 2 \left\| \mu_{\mathcal{D}} \right\| \left| v \cdot \left(\mu_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} x \right) \right| . \end{aligned}$$

Lemma 61 implies that the variable $(v \cdot x)^2$ for $x \sim \mathcal{D}$, scaled by a factor of

$$\frac{1}{(2 + \alpha + \beta)^{O_m(1)}} ,$$

also satisfies the conditions of Claim 28 with $c = \Omega_m(1)$. Also note that

$$\mathbb{E}_{x \sim \mathcal{D}} (v \cdot x)^2 = v^T (\Sigma_{\mathcal{D}} + \mu_{\mathcal{D}} \mu_{\mathcal{D}}^T) v .$$

Thus, we can ensure that with $1 - e^{(8\kappa/\epsilon)^2}$ probability, we have

$$\left| v^T (\Sigma_{\mathcal{D}} + \mu_{\mathcal{D}} \mu_{\mathcal{D}}^T) v - \frac{1}{S'} \sum_{x \in S'} (v \cdot x)^2 \right| \leq \frac{\delta^2}{10\epsilon} ,$$

for all subsets S' with $|S'| \geq (1-\epsilon)n$. Lemma 61 also implies that

$$\| \mu_{\mathcal{D}} \| \leq (2 + \alpha + \beta)^{O_m(1)}$$

so overall, we conclude that with $1 - e^{(7\kappa/\epsilon)^2}$ probability, for each fixed vector v, we have both

$$\left| v \cdot \left(\mu_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} x \right) \right| \leq \frac{\delta}{2}$$

$$\left| v^T \left(\Sigma_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} (x - \mu_{\mathcal{D}})(x - \mu_{\mathcal{D}})^T \right) v \right| \leq \frac{\delta^2}{2\epsilon}$$

for all subsets S' with $|S'| \geq (1-\epsilon)n$. A union bound gives that with $1 - e^{(6\kappa/\epsilon)^2}$ probability, the above holds for all vectors $v \in E$. Now let b be the vector

$$b = \left(\mu_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} x \right)$$

and let Q be the matrix

$$Q = \left(\Sigma_{\mathcal{D}} - \frac{1}{S'} \sum_{x \in S'} (x - \mu_{\mathcal{D}})(x - \mu_{\mathcal{D}})^T \right) .$$

We want to bound the L^2 norm of b and the operator norm of Q. First, let v be the element of E that is closest to the unit vector in the direction of b. Since E is a $\epsilon/(10\kappa)$-net, we must have $v \cdot b > 0.5 \left\| b \right\|_2$ which implies $\left\| b \right\|_2 \leq \delta$. Next let u be a unit vector such that $u^T Q u = \left\| Q \right\|_{\mathsf{op}}$ (such a u exists since Q is symmetric). Let v be the element of E closest to u. Then

$$\begin{aligned} v^T Q v = u^T Q u + (v-u)^T Q u + v^T Q (v-u) &\geq \left\| Q \right\|_{\mathsf{op}} - |(v-u)^T Q u| - |v^T Q (v-u)| \\ &\geq \left\| Q \right\|_{\mathsf{op}} (1 - 2 \left\| v - u \right\|_2) > \frac{\left\| Q \right\|_{\mathsf{op}}}{2} \end{aligned}$$

and thus we must actually have $\left\| Q \right\|_{\mathsf{op}} \leq \delta^2/\epsilon$. This completes the proof. ■