CS 224 Fall 2023

Yuyang Zhang

Difficulties & Solutions in Learning Latent Variable Models

1 Project Topic Overview

This report studies the problem of learning high-dimensional models with hidden variables, with a special focus on *Gaussian Mixture Models (GMM)*. The goal is to present an algorithm framework that achieves quasi-polynomial dependence on the hidden parameter size k, and to articulate the key insight as compared to other literature that depends exponentially on k.

We will first review previous algorithms for GMM [SOAJ14] and explain intuitively their inherent exponential dependence on the hidden parameter k. We then present the algorithm framework proposed in this work [DK20]. As compared to the original paper, we largely reorganize the flow so that more focus is put on the intuitions behind this method and connections to previous algorithms, instead of including all the technical details. We also plan to carefully present the core of their proof and highlight the key component that enables the quasi-polynomial dependence.

The rest of the report is organized as follows. In Section 2, we formally introduce the GMM model and learning target. In Section 3, we introduce and rephrase a family of previous algorithms for GMM. In Section 4, the quasi-polynomial algorithm framework is introduced. We decompose it into four steps, explain in detail how every step is executed, and prove the efficiency and correctness. In Section 5, we specifically focus on step 3, the construction of the small cover, of the aforementioned algorithm framework. We discuss in great detail the reasons and intuitions behind the recursive cover construction. Core proofs are presented. Finally, Section 6 summarizes the report and points out several future directions.

1.1 Preliminary

For any integer k, let $[k] = \{1, \ldots, k\}$. For any polynomial p, let $\|p\|$ be its l_2-norm, and let A_p be the tensor s.t. $p(x) = \langle A_p, x^{\otimes \text{degree}(p)} \rangle$. $\mathbb{R}^m_{[d]}$ space of degree-d polynomials defined on $\mathbb{R}^m$. For any vector v, let $\|v\|$ be its 2-norm. For any linear subspace, let $\mathcal{V}^\perp$ be its orthogonal complement. Let co-dim be the abbreviation for co-dimension. We use $\tilde{\mathcal{O}}$ to hide logarithmic factors and constants. For any polynomial space $\mathcal{P}$, we also use "variety" to denote the set of points where $\forall p \in \mathcal{P}$ nearly vanishes (formally defined in Equation 33).

2 Problem Formulation

Throughout the first few sections, we will use the Gaussian mixture model (GMM) as a motivating example. In fact, the methods will be extendable to other models, including RELU networks, mixture of linear regressions, etc.

Definition 1 (Gaussian Mixture Model (GMM))**.** A Gaussian mixture model refers to the following probability distribution, which is a k-Mixture of m-Dimensional Spherical Gaussians

$$x \sim F(x) \coloneqq \sum_{i \in [k]} w_i \mathcal{N}(v_i, I_d), \tag{1}$$

where $\|v\|_i \leq R$, $w_i > 0$ for some universal constant R and $\forall i \in [k]$. Moreover, $\sum_{i \in [k]} w_i = 1$. The parameters are summarized as follows

$$k, m, R, \{w_i\}_{i \in [k]}, \{v_i\}_{i \in [k]}. \tag{2}$$

For the rest of the report, we use polynomial to refer to polynomial in k, m, R and ϵ, where ϵ is some small constant indicating the accuracy of the algorithm. We use quasi-polynomial to refer to quasi-polynomial in k and polynomial in m, R, ϵ.

We first sample a dataset $\mathcal{D} = \{x_i\}_{i \in [N]}$ from distribution F, where N is the count of samples. Utilizing this dataset $\mathcal{D}$, our target is density estimation, which we will abbreviate as "learning" for simplicity. Namely, we want to output a hypothesis distribution $\widehat{F}(\mathcal{D})$ such that the following holds with high probability

$$d_{\text{TV}}(\widehat{F}(\mathcal{D}), F) \leq \tilde{\mathcal{O}}(\epsilon). \tag{3}$$

3 Previous Learning Algorithms

3.1 Direct Subspace Learning

Let's first start with the following naive attempt. Let $\mathcal{V}$ be the space spanned by $\{v_i\}_{i \in [k]}$ with dimension $k^- \leq k$. From dataset $\mathcal{D}$, we can easily estimate the second moment M_2 of the parameters $\{v_i\}_{i \in [k]}$

$$M_2 = \sum_{i \in [k]} w_i v_i v_i^\top = \sum_{i \in [k]} w_i v_i^{\otimes 2}. \tag{4}$$

Solving for the column space of this matrix directly gives us the space $\mathcal{V}$.

Now we construct an ϵ-cover of $\mathcal{V}$, denoted by $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$. Notice that all parameters $\{v_i\}_{i\in[k]}$ lie in $\mathcal{V}$, by definition. Therefore, for any $i \in [k]$, there exist a $j(i) \in [|\mathcal{C}|]$ such that

$$\left\| c_{j(i)} - v_i \right\| \le \epsilon. \tag{5}$$

Based on this intuition, we construct the following hypothesis $\widehat{F}$

$$\widehat{F} = \sum_{j=1}^{|\mathcal{C}|} \widehat{w}_j \mathcal{N}(c_j, I), \tag{6}$$

and solve for the optimal distribution by some convex optimization techniques. Under the optimal $\{\widehat{w}_j^*\}_{j=1}^{|\mathcal{C}|}$, the $\widehat{F}^*$ should perform better than the following distribution

$$\sum_{i=1}^{k} w_i \mathcal{N}(c_{j(i)}, I). \tag{7}$$

Since the above distribution is already very close to the original distribution F (because $c_{j(i)}$ is very close to v_i), our solution $\widehat{F}^*$ should be even closer and therefore satisfies our requirement.

A lot of previous works developed algorithms based on this simple intuition. However, this type of method suffers from an inherent difficulty. The ϵ-cover $\mathcal{C}$ that we constructed is of size $\mathcal{O}(2^k)$, since it covers the k-dimensional subspace $\mathcal{V}$. Therefore, our hypothesis is composed of $\mathcal{O}(2^k)$ terms, and the convex optimization problem takes time exponential in k.

3.2 Getting Rid of The Exponential Dependence on k

Previous work [DKS17] managed to show a lower bound of $m^{\Omega(k)}$ on the complexity of learning k-mixture of m-dimensional Gaussians. However, their lower bound construction highly relies on non-spherical Gaussian distributions, but here we are only considering spherical Gaussians. Namely, it is still possible to get subexponential dependence on k.

The most straightforward idea is to adapt existing algorithms in related fields to our setting. There actually exists polynomial time and sample algorithms based on list-decodable learning of Gaussian distributions [DKS18, HL18, KSS18]. However, all of these algorithms require the individual Gaussian distributions to be well separated, which is not the case for our GMM.

Another idea is trying to construct smaller, i.e. subexponential-size, covers for the set of possible mean vectors, which actually works to provide an algorithm with quasi-polynomial time and sample complexity with the techniques in the following sections.

3.3 Linear Variety Learning

Before going any further, we will first recast the aforementioned "direct subspace learning" method (Section 3.1) in a more general way. We can actually think of it as an algorithm learning a linear variety. Following this slightly different formulation, this simple algorithm can then be extended towards quasi-polynomial time and sample complexity.

The main steps of the "direct subspace learning" are summarized as follows, assume we have access to the exact second moment $M_2 = \sum_{i\in[k]} w_i v_i^{\otimes 2}$:

- Solve for the null space of M_2, i.e. $\mathcal{V}^\perp$.
- Construct the ϵ-cover for $(\mathcal{V}^\perp)^\perp = \mathcal{V}$, denote as $\mathcal{C} = \{c_i\}_{i=1}^{|\mathcal{C}|}$.
- Form hypothesis $\widehat{F} = \sum_{i=1}^{|\mathcal{C}|} \widehat{w}_i \mathcal{N}(c_i, I)$ and solve for the optimal $\{\widehat{w}_i^*\}_{i=1}^{|\mathcal{C}|}$.

Here one may notice that the first and second lines can be combined as one step: constructing an ϵ-cover for the column space of M_2. Here we rephrase it in this "tilted way" so that more intuitions can be drawn afterwards.

We now analyze the first step, which calculates an $(m - k^-)$-dimensional subspace orthogonal to $\mathcal{V}$, i.e. $\mathcal{V}^\perp$. Recall that there is a bijection between all vectors in $\mathbb{R}^m$ and all degree-1 polynomials in $\mathbb{R}^m_{[1]}$. To be more specific, for any vector $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_m)^\top \in \mathbb{R}^m$, the corresponding polynomial is

$$p_{\tilde{v}}(x) = \sum_{i\in[m]} \tilde{v}_i x_i. \tag{8}$$

Now for $\forall c \in \mathcal{V}^\perp$, we know that

$$p_c(v) = \sum_{j\in[m]} c_j v_j = c^\top v = 0, \quad \forall v \in \mathcal{V}. \tag{9}$$

Thus, $p_c(x)$ vanishes at all $v \in \mathcal{V}$ and therefore vanishes at the k points $\{v_i\}_{i\in[k]}$. Denote this set of polynomials as $\mathcal{P} = \{p_c(x), \forall c \in \mathcal{V}^\perp\}$. Moreover, we can conclude that $\mathcal{P}$ has dimension $m - k^-$, or equivalently, has co-dim k^-. This is because the dimension of $\mathcal{V}^\perp$ is $m - k^-$, and because the dimension of $\mathcal{P}$ equals the dimension of $\mathcal{V}^\perp$ (which follows naturally from Equation 8). Therefore, the first step is rephrased:

Solving for the subspace of degree-1 polynomials with co-dim k^- that vanishes on $\{v_i\}_{i\in[k]}$.

We slightly overload the notation by letting the solution to the above step be $\mathcal{P}$. To rephrase the second step, we first prove that the variety of $\mathcal{P}$ is just the space $\mathcal{V}$. Firstly, the variety includes $\mathcal{V}$. This is because for any $v = \sum_{i\in[k]} \alpha_i v_i \in \mathcal{V}$, we have

$$p_c(v) = c^\top v = \sum_{i\in[k]} \alpha_i c^\top v_i = 0, \quad \forall p_c \in \mathcal{P}. \tag{10}$$

On the other hand, the dimension of the variety is k^-. To see this, notice that the dimension of $\mathcal{P}$ equals the dimension of the following subspace

$$\{c \in \mathbb{R}^m : p_c \in \mathcal{P}\}, \tag{11}$$

which is $m - k^-$. Moreover, notice that the variety of $\mathcal{P}$ is just the orthogonal complement of the above subspace, which therefore has dimension $m - (m - k^-) = k^-$. Combining the two parts shows that the variety of $\mathcal{P}$ is $\mathcal{V}$. Since the second step is to cover the space $\mathcal{V}$, we naturally rephrase it as the following based on the above intuition

Construct an ϵ-cover for the variety of $\mathcal{P}$.

Denote the cover by $\mathcal{C}$. Notice that points in cover $\mathcal{C}$ should approximate parameters $\{v_i\}_{i\in[k]}$ well, since they cover the whole space $\mathcal{V}$. Therefore, we can use $\mathcal{C}$ to form our hypothesis.

Finally, we list the rephrased steps as follows, assuming we have access to the exact second moment $M_2 = \sum_{i\in[k]} w_i v_i^{\otimes 2}$:

- From M_2, solve for $\mathcal{P} \subset \mathbb{R}^m_{[1]}$ that vanishes on $\{v_i\}_{i\in[k]}$ with dimension $m - k^-$.
- Construct an ϵ-cover for the variety of $\mathcal{P}$, denoted by $\mathcal{C} = \{c_i\}_{i=1}^{|\mathcal{C}|}$.
- Form hypothesis $\widehat{F} = \sum_{i=1}^{|\mathcal{C}|} \widehat{w}_i \mathcal{N}(c_i, I)$ and solve for optimal parameters $\{\widehat{w}_i^*\}_{i=1}^{|\mathcal{C}|}$.

At this point, our objective is to construct a smaller ϵ-cover that covers the set of points $\{v_i\}_{i\in[k]}$. Notice that we get this cover from the variety of the polynomial space $\mathcal{P}$. A natural idea is to construct a more complicated polynomial space $\mathcal{P}$ that vanishes on $\{v_i\}_{i\in[k]}$. Intuitively, more complicated $\mathcal{P}$ leads to smaller variety. And therefore fewer points are needed to cover this variety.

In the following sections, we prove that this idea actually works and leads to algorithms with quasi-polynomial time and sample complexity.

4 The "Small Covers" Method

4.1 The Algorithm

It turns out that to get a small cover, it suffices to consider $\mathcal{P} \subset \mathbb{R}^m_{[d]}$, i.e. degree-d polynomials on $\mathbb{R}^m$, where d is chosen to be $\log(k)$. This gives the following high-level procedure, when we are given the exact $2d$-th moment $M_d = \sum_{i\in[k]} w_i v_i^{\otimes 2d}$:

- From M_d, get $\mathcal{P} \subset \mathbb{R}^m_{[d]}$ vanishing on $\{v_i\}_{i\in[k]}$ with co-dim at most k.
- Construct an ϵ-cover for the variety of $\mathcal{P}$, denoted by $\mathcal{C} = \{c_i\}_{i=1}^{|\mathcal{C}|}$.
- Form hypothesis $\widehat{F} = \sum_{i=1}^{|\mathcal{C}|} \widehat{w}_i \mathcal{N}(c_i, I)$ and solve for optimal parameters $\{\widehat{w}_i^*\}_{i=1}^{|\mathcal{C}|}$.

In real applications where M_d is not directly available, the above algorithm is directly extended with approximation techniques to be mentioned later. We then have the following approximation algorithm

1. Approximate the d-th moment $M_d \approx \sum_{i\in[k]} w_i v_i^{\otimes 2d}$.
2. From M_d, get $\mathcal{P} \subset \mathbb{R}^m_{[d]}$ *nearly vanishing* on $\{v_i\}_{i\in[k]}$ with co-dim at most k.
3. Construct an ϵ-cover for the set of points where $\mathcal{P}$ *nearly vanishes*, denoted by $\mathcal{C} = \{c_i\}_{i=1}^{|\mathcal{C}|}$.
4. Form hypothesis $\widehat{F} = \sum_{i=1}^{|\mathcal{C}|} \widehat{w}_i \mathcal{N}(c_i, I)$ and solve for optimal parameters $\{\widehat{w}_i^*\}_{i=1}^{|\mathcal{C}|}$.

For the rest of this section, we will give detailed explanations of how each step is executed. Then we prove:

1. Every step can be finished within quasi-polynomial time.
2. Results of every step satisfy the corresponding requirements.

Some core proofs will be deferred to the next section.

4.2 Step 1: Estimating Moments

Let $H_n(t), t \in \mathbb{R}, n \in N$ denote the following probabilist Hermite polynomial

$$H_n(t) = (-1)^n e^{t^2} \frac{\mathrm{d}^n}{\mathrm{d}t^n} e^{-t^2}. \tag{12}$$

We set the a-th entry of our moment estimation M_d to be

$$\frac{1}{N}\sum_{n=1}^{N}\left(\prod_{j=1}^{m} H_{a_j}(x_j^n)\right) \tag{13}$$

for every index tuple $a = (a_1, \dots, a_m) \in \mathbb{N}^m$ with $\|a\|_1 = 2d$.

The construction comes from the following observation

$$\sum_{i\in[k]} w_i v_i^a = \mathbb{E}_{x=(x_1,\dots,x_m)}\left(\prod_{j=1}^{m} H_{a_j}(x_j)\right). \tag{14}$$

Here $v_i^a = \prod_{j=1}^{m} v_{i,j}^{a_j}$ and $x \sim F$ follows the GMM defined in Equation 1. Intuitively, this means that a-th entry of the tensor $\sum_{i\in[k]} w_i v_i^{\otimes 2d}$ can be well approximated by averaging $\prod_{j=1}^{m} H_{a_j}(x_j)$ of all samples. Then combining them together should result in a good enough moment estimation.

To be more specific, with $N = (Rmd)^{\mathcal{O}(d)}/\delta^2$ samples from F, we can make sure that the following holds

$$\left|\sum_{i\in[k]} w_i v_i^a - \frac{1}{N}\sum_{n=1}^{N}\left(\prod_{j=1}^{m} H_{a_j}(x_j^n)\right)\right| \le \delta^2 \tag{15}$$

Therefore, as long as we have enough samples from distribution F, all entries of the tensor can be well approximated.

Moreover, the estimations of all entries can be finished in quasi-polynomial time. This is because we have at most $\mathcal{O}(m^{2d}) = \mathcal{O}(m^{\mathcal{O}(\log(k))})$ entries for the tensor, and the estimation of every single entry takes polynomial time.

To conclude this subsection, we make the above claims formal by the following lemma. We omit its proof because it is not closely related to the core of the "small cover" method.

Lemma 1. With $N = (Rmd)^{\mathcal{O}(d)}/\delta^2$ samples from F, we can compute a tensor M_d with the a-th entry being $\frac{1}{N}\sum_{n=1}^{N}\left(\prod_{j=1}^{m} H_{a_j}(x_j^n)\right)$ that satisfies

$$\left\|M_d - \sum_{i\in[k]} w_i v_i^{\otimes 2d}\right\| \le \mathcal{O}(\delta) \tag{16}$$

in time $poly(N)$, i.e. in quasi-polynomial time.

4.3 Step 2: Approximating Polynomial Space $\mathcal{P}$

Based on M_d computed in the previous step, we construct $\mathcal{P}$ as follows. Consider the following map $Q : \mathbb{R}^m_{[d]} \to \mathbb{R}$:

$$Q(p) = \langle A_{p^2}, M_d \rangle. \tag{17}$$

Let $\mathcal{P} \subset \mathbb{R}^m_{[d]}$ be the subspace spanned by **all but the top-k eigenvectors of** $Q(p)$.

This construction is pretty intuitive. By excluding the top-k eigenvector of $Q(p)$, we are left with a space of polynomials where $Q(p)$ is very small. Notice that $Q(p) \approx \sum_{i\in[k]} w_i p^2(v_i)$. Then it follows naturally that this space should nearly vanish on $\{v_i\}_{i\in[k]}$. We then prove its efficiency and that it satisfies all requirements in Step 2.

By construction, we know that $\mathcal{P}$ has codimension at most k. Moreover, we can construct $\mathcal{P}$ in quasi-polynomial time. To see this, consider any degree-d polynomial $p = \sum_i \alpha_p^i p_i$, where $\{p_i\}$ is a basis for $\mathbb{R}^m_{[d]}$ with size $\binom{m+d-1}{d}$ and $\alpha_p := (\alpha_p^1, \dots, \alpha_p^{\binom{m+d-1}{d}})$ is the coefficient. Then $Q(p)$ is just the quadratic form of α_p. This implies that $\mathcal{P}$ can be solved by solving for the top-k eigenvectors of this quadratic form, which can be solved within time polynomial in the dimension of α_p. This dimension is upper bounded by

$$\binom{m+d-1}{d} \le (m+d)^d = \mathcal{O}(m)^{\log(k)} \tag{18}$$

due to our choice of $d = \log(k)$. Therefore, solving for $\mathcal{P}$ takes quasi-polynomial time.

For the rest of this section, we prove $\mathcal{P}$ nearly vanishes on $\{v_i\}_{i\in[k]}$.

Lemma 2. Given an accurate enough moment estimation M_d satisfying

$$\left\| M_d - \sum_{i\in[k]} w_i v_i^{\otimes 2d} \right\| \le \delta, \tag{19}$$

$\mathcal{P}$ *nearly vanishes* on $\{v_i\}_{i\in[k]}$ in the following sense

$$|p(v_i)| \le \mathcal{O}(\sqrt{\frac{\delta}{w_i}}\, \|p\|), \quad \forall i \in [k], \forall p \in \mathcal{P}. \tag{20}$$

Proof. To finish the proof, we need the following simple claim

$$\left\| A_{p^2} \right\| \le \|p\|^2 . \tag{21}$$

Now we prove that $\mathcal{P}$ satisfies our requirements (Equation 20) by contradiction.

Notice that for any $p \in \mathcal{P}$,

$$\begin{aligned} Q(p) &= \langle A_{p^2}, M_d \rangle = \langle A_{p^2}, \sum_{i\in[k]} w_i v_i^{\otimes 2d} \rangle + \langle A_{p^2}, M_d - \sum_{i\in[k]} w_i v_i^{\otimes 2d} \rangle \\ &= \sum_{i\in[k]} w_i \langle A_{p^2}, v_i^{\otimes 2d} \rangle + \langle A_{p^2}, M_d - \sum_{i\in[k]} w_i v_i^{\otimes 2d} \rangle \\ &= \sum_{i\in[k]} w_i p^2(v_i) + \mathcal{O}(\left\|A_{p^2}\right\| \left\|M_d - \sum_{i\in[k]} w_i v_i^{\otimes 2d}\right\|) \\ &= \sum_{i\in[k]} w_i p^2(v_i) + \mathcal{O}(\delta \left\|p\right\|^2). \end{aligned} \tag{22}$$

Suppose $\exists p \in \mathcal{P}$ such that $Q(p) = \Omega(\delta \left\|p\right\|^2)$. Then for any p' in the space spanned by the top-k eigenvectors of Q, the following holds

$$Q(p') = \Omega(\delta \left\|p'\right\|^2). \tag{23}$$

The existence of such p, together with the top-k eigenvectors of $Q(p)$, implies that the following space has co-dim at least $k+1$:

$$\mathcal{S} = \left\{ p \in \mathbb{R}^m_{[d]} : \sum_{i\in[k]} w_i p^2(v_i) = 0 \right\}. \tag{24}$$

However, the co-dim of the above space is upper bounded by k, since there are at most k constraints on this space, specified by every $v \in \{v_i\}_{i\in[k]}$. This leads to a contradiction. Therefore, the following holds for $\forall p \in \mathcal{P}$

$$Q(p) = \sum_{i\in[k]} w_i p^2(v_i) = \mathcal{O}(\delta \left\|p\right\|^2) \geq w_i p^2(v_i), \quad \forall i \in [k]. \tag{25}$$

which translates to

$$p(v_i) \leq \mathcal{O}(\sqrt{\frac{\delta}{w_i}} \left\|p\right\|), \quad \forall i \in [k], \forall p \in \mathcal{P}. \tag{26}$$

□

4.4 Step 3: Constructing ϵ-Cover for The "Variety" of $\mathcal{P}$

Let C denote some large enough constant. For this subsection, fix a small constant ϵ and choose $\delta = \epsilon^{2d} \left(\frac{\epsilon}{2Rkdm}\right)^{4Cd}$. Now we have polynomial space $\mathcal{P}$ such that the following hold for all $p \in \mathcal{P}$

$$p(v_i) \leq \mathcal{O}(\sqrt{\frac{\delta}{w_i}}\,\|p\|). \tag{27}$$

Therefore, for all $w_i \geq \epsilon/k$, we know that

$$|p(v_i)| \leq \mathcal{O}\left(\epsilon^d \left(\frac{\epsilon}{2Rkdm}\right)^{2Cd} \sqrt{\frac{k}{\epsilon}}\,\|p\|\right) \leq \mathcal{O}\left(\epsilon^d \left(\frac{\epsilon}{2Rkdm}\right)^{Cd} \|p\|\right). \tag{28}$$

Let $\mathcal{I} = \{i : w_i \geq \epsilon/k\}$. We now try to find an ϵ-cover with quasi-polynomial size, i.e. a small cover, for the following set $\mathcal{S}$

$$\mathcal{S} = \left\{x \in \mathbb{R}^m : \|x\| \leq R \text{ and } |p(x)| \leq \epsilon^d \left(\frac{\epsilon}{2Rdkm}\right)^{Cd} \|p\|, \forall p \in \mathcal{P}\right\}. \tag{29}$$

This is the subspace of $\mathbb{R}^m$ where all polynomials in $\mathcal{P}$ *nearly vanish*, which clearly includes our parameters $\{v_i\}_{i\in[k]\setminus\mathcal{I}}$ according to the construction of $\mathcal{P}$. Covering this subspace therefore gives us good approximations about those parameters. Other parameters $\{v_i\}_{i\in\mathcal{I}}$ are not so important, because their coefficients are too small.

For simplicity, we defer the detailed construction of such small ϵ-covers to the next section. Instead, in this subsection, we talk about the intuitions behind the reason why there should exist a small cover with only quasi-polynomial size.

To see this, let's consider a special case where $\delta = 0$. Namely, we are given polynomial space $\mathcal{P}^* = \{p \in \mathbb{R}^m_{[d]} : p(v_i) = 0,\ \forall i \in [k]\}$. Similarly, define $\mathcal{S}^*$ to be the following space where all polynomials in $\mathcal{P}^*$ exactly vanish

$$\mathcal{S}^* = \left\{x \in \mathbb{R}^m : \|x\| \leq R \text{ and } |p(x)| = 0, \forall p \in \mathcal{P}^*\right\}. \tag{30}$$

Then it is clear that $\mathcal{S}^*$ is just the variety of $\mathcal{P}^*$ in the usual sense. Let $\mathcal{Q}^*$ be the space of degree-d polynomials defined on $\mathcal{S}^*$. From basic algebraic geometry, we know that $\mathcal{Q}^*$ is isomorphic to $\mathbb{R}^m_{[d]}/\mathcal{P}^*$. Since $\mathcal{P}^*$ has codimension at most k by construction, we know that $\mathbb{R}^m_{[d]}/\mathcal{P}$, and therefore $\mathcal{Q}^*$, has dimension at most k.

This fact is clearer when we consider the special case when $d = 1$ and the co-dimension of $\mathcal{P}$ is at most k. Due to the bijection between all degree-1 one polynomials $p(x) = \sum_{j\in[m]} \alpha_x^j x_j$ and vectors $\alpha_x = (\alpha_x^1, \ldots, \alpha_x^m)$, space $\mathcal{P}$ defines an at least $(m-k)$-dimensional subspace of $\mathbb{R}^m$, denoted by $\mathcal{V}^\perp$. Meanwhile, the variety of $\mathcal{P}$ is all the vector v satisfying $\sum_{j\in[m]} \alpha_x^j v_j = \alpha_x^\top v = 0$. Therefore, the variety of $\mathcal{P}$ is just the orthogonal complement of $\mathcal{V}^\perp$, i.e. $\mathcal{V}$. It is clear that, from a linear algebra perspective, $\mathcal{V}$ has dimension at most k. And again, due to the bijection between degree-1 polynomials and vectors, the space of degree-1 polynomials defined on $\mathcal{V}$ is defined as follows

$$\left\{ p : p = \sum_{j \in [m]} \alpha_j x_j, \quad \forall \alpha \in \mathcal{V} \right\}, \tag{31}$$

which has dimension at most k.

Now that the fact $\dim(\mathcal{Q}^*) \leq k$ is clear, we use it to upper bound $\dim(\mathcal{S}^*)$. Since the space of all degree-d polynomials on $\mathcal{S}^*$ has dimension $\binom{\dim(\mathcal{S}^*)+d-1}{d}$, we know

$$\binom{\dim(\mathcal{S}^*) + d - 1}{d} \leq k. \tag{32}$$

This implies $\dim(\mathcal{S}^*) = \mathcal{O}(dk^{1/d})$. Therefore, to cover the space $\mathcal{S}^*$, one may expect the size of the cover to be exponential in $\mathcal{O}(dk^{1/d})$. By letting $d = \log(k)$, we get a quasi-polynomial-sized cover.

The following lemma formalizes all the above intuitions, whose proof is deferred to the next section.

Lemma 3. Let C be some large enough constant. Fix a small constant ϵ. Given $\mathcal{P}$ from step 2 with $\delta = \epsilon^{2d} \left(\frac{\epsilon}{2Rkdm}\right)^{4Cd}$. Consider the following space $\mathcal{S}$

$$\mathcal{S} = \left\{ x \in \mathbb{R}^m : \|x\| \leq R \;\text{ and }\; |p(x)| \leq \epsilon^d \left(\frac{\epsilon}{2Rdkm}\right)^{Cd} \|p\|, \forall p \in \mathcal{P} \right\}. \tag{33}$$

There exists an ϵ-cover of $\mathcal{S}$ with size at most $(2Rdkm/\epsilon)^{C^2 d^2 k^{1/d}}$.

4.5 Step 4: Optimizing Hypothesis $\widehat{F}$

With the quasi-polynomial-sized cover $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ from previous subsection, we form our hypothesis as follows

$$\widehat{F} = \sum_{j \in |\mathcal{C}|} \widehat{w}_j \mathcal{N}(c_j, I), \tag{34}$$

Since $\mathcal{C}$ covers the variety of $\mathcal{P}$, which includes $\{v_i\}_{i\in[k]}$, we know that $\widehat{F}$ approximates F well for some coefficients $\{\widehat{w}_i\}_{i\in[|\mathcal{C}|]}$. Denote the set of all possible $\widehat{F}$'s with coefficients "not too small" (which will be formally defined later) as Δ, which is a convex set. We then try to optimize over Δ for the best possible distribution.

We first make the following definitions for any distribution f with the same support as F and sample $x \sim F$

$$L(f, x) := \log(f(x)), \quad L(f) := \mathbb{E}_{x\sim F}[L(f, x)] = D(F||f) + H(F). \tag{35}$$

Here $D(F||f)$ denotes the KL-divergence between F and f, and $H(F)$ denotes the entropy. Our final hypothesis is then chosen to be the minimizer of the following optimization problem

$$\widehat{F} = \min_{f\in\Delta} \frac{1}{N}\sum_{n=1}^{N} L\left(f, x_n\right). \tag{36}$$

Now we give intuition behind the above optimization problem. From definition, we know that $\frac{1}{N}\sum_{n=1}^{N} L\left(f, x_n\right)$ is the empirical estimator of $L(f)$. **Thanks to the low-dimensionality of Δ (quasi-polynomial), we know that with quasi-polynomially many samples, $\frac{1}{N}\sum_{n\in[N]} L(f, x_i)$ approximates $L(f)$ well for every single $f \in \Delta$.** On the other hand, if f and F are close, the $D(F||f)$ should be small. And therefore $L(f)$ should also be small. With good estimates of $L(f)$, we can directly choose the best f that minimizes $L(f)$ as our estimation $\widehat{F}$, which naturally leads to a good approximation of the original distribution.

We now formally prove its efficiency and correctness. From previous literature [DSS18], we know that this problem can be solved up to error ϵ within polynomial time. The correctness is established by the following Lemma

Lemma 4. Given a quasi-polynomial-sized cover $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ that satisfies the following for $\forall i \in [k]\backslash\mathcal{I}$

$$\exists j, \text{ s.t. } \left\|c_j - v_i\right\| \leq \epsilon. \tag{37}$$

Recall here $\mathcal{I} = \{i : w_i \geq \epsilon/k\}$. Then with $N = \mathcal{O}(|\mathcal{C}|/\epsilon^2)$ samples, i.e. quasi-polynomially many samples, we can calculate a distribution $\widehat{F}$ within polynomial time such that

$$d_{\text{TV}}(\widehat{F}, F) \leq \mathcal{O}(\sqrt{\epsilon\log(\frac{|\mathcal{C}|}{\epsilon})}). \tag{38}$$

Proof. Let $\Delta = \{f : f = \sum_{j\in|\mathcal{C}|} w_j\mathcal{N}(c_j, I), \sum_{j\in|\mathcal{C}|} w_j = 1, \boldsymbol{w_j > \epsilon/k}, \forall\{w_j\}_{j\in[|\mathcal{C}|]}\}$, which contains all possible Gaussian mixtures constructed from cover $\mathcal{C}$ with coefficients not too small.

For simplicity of the proof, we claim without proving the following fact. For any pair $p, q \in \Delta$, one necessary condition of the following equality

$$\begin{aligned}\widehat{L}(p) - \widehat{L}(q) &\coloneqq \frac{1}{N}\sum_{n=1}^{N} L\left(p, x_n\right) - \frac{1}{N}\sum_{n=1}^{N} L\left(q, x_n\right) \\ &= L(p) - L(q) + \mathcal{O}(\epsilon\log(|\mathcal{C}|/\epsilon))\end{aligned} \tag{39}$$

is that the following quantity is **at most quasi-polynomial** for all $\{a_i\}_{i\in[|\mathcal{C}|]}$

$$\mathrm{VC}\left(\left\{\sum_{i\in[|\mathcal{C}|]} a_i p_i(x), \forall x \in \mathbb{R}^m\right\}\right). \tag{40}$$

Here VC($\mathcal{A}$) denotes the VC-dimension of set $\mathcal{A}$, p_i denotes the distribution $\mathcal{N}(c_i, I)$. Intuitively, to accurately approximate $\widehat{L}(p) - \widehat{L}(q)$ for every pair $p, q \in \Delta$ with only quasi-polynomial number of samples, we need space Δ to not too complicated. The "complexity" of Δ is measured by the VC-dimension of the linear combination of its basis $\{p_i, i \in [|\mathcal{C}|]\}$. As long as this complexity is quasi-polynomial, quasi-polynomially many samples will be enough to approximate $L(p) - L(q)$ for every pair $p, q \in \Delta$[1].

In fact, the above VC-dimension equals $\mathcal{O}(|\mathcal{C}|)$. To see this, let's do the following change of variable: $x \to (p_1(x), \ldots, p_{|\mathcal{C}|}(x))$. Then the above set becomes the set of all halfspaces in $\mathbb{R}^{|\mathcal{C}|}$, which has VC-dimension $\mathcal{O}(|\mathcal{C}|)$. In our case, this quantity is quasi-polynomial. Therefore, Equation 39 holds.

For the rest of this subsection, we will use Equation 39 to finish the proof of this Lemma. We first show that there exists a distribution f in Δ such that

$$d_{\mathrm{TV}}(f, F) \leq 2\epsilon. \tag{41}$$

Choose $\{\widehat{w}_j\}_{j\in[|\mathcal{C}|]}$ as follows. For all $i \in [k]$ with $w_i > \epsilon/|\mathcal{C}|$, set $\widehat{w}_j = w_i$ for any one of the j's satisfying $\left\|c_j - v_i\right\| \leq \epsilon$. Denote this index j by $j(i)$. Set all other $\widehat{w}_j$'s to zero. Denote the distribution with this set of coefficients as f^*. For all $i \in [k]$ with

[1]Formal proof of this claim is in the proof of Proposition 28 in [DK20]

$w_i \leq \epsilon/|\mathcal{C}|$, let them form set $\mathcal{I}$. Then we have

$$
\begin{aligned}
d_{\mathrm{TV}}(f^*, F) &= d_{\mathrm{TV}}\left(\sum_{j\in|\mathcal{C}|} w_j \mathcal{N}(c_j, I), F\right) \\
&= d_{\mathrm{TV}}\left(\sum_{i\in[k]\setminus\mathcal{I}} w_i \mathcal{N}(c_{j(i)}, I), \sum_{i\in[k]} w_i \mathcal{N}(v_i, I)\right) \\
&\leq \sum_{i\in[k]\setminus\mathcal{I}} w_i d_{\mathrm{TV}}\left(\mathcal{N}(c_{j(i)}, I), \mathcal{N}(v_i, I)\right) + \sum_{i\in\mathcal{I}} w_i \\
&\overset{(i)}{\leq} \sum_{i\in[k]\setminus\mathcal{I}} w_i \sqrt{\frac{1}{2} d_{\mathrm{KL}}(\mathcal{N}(c_{j(i)}, I), \mathcal{N}(v_i, I))} + \epsilon \\
&= \sum_{i\in[k]\setminus\mathcal{I}} w_i \sqrt{\frac{1}{2}\left\|c_{j(i)} - v_i\right\|^2} + \epsilon \\
&\leq \sum_{i\in[k]\setminus\mathcal{I}} w_i \epsilon + \epsilon \leq 2\epsilon.
\end{aligned}
\tag{42}
$$

Here (i) is from Pinsker's inequality.

Finally, we show the output of our optimization problem (Equation 36), denoted by $\widehat{f}$, is close enough to f^*. In polynomial time, we have the following for $\forall f \in \Delta$

$$
\widehat{L}(\widehat{f}) \leq \widehat{L}(f) + \epsilon \iff \widehat{L}(\widehat{f}) - \widehat{L}(f) \leq \epsilon. \tag{43}
$$

Take $f = f^*$, and from Equation 39, we have

$$
L(\widehat{f}) - L(f^*) \leq \widehat{L}(\widehat{f}) - \widehat{L}(f^*) + \mathcal{O}(\epsilon \log(|\mathcal{C}|/\epsilon)) \leq \mathcal{O}(\epsilon \log(|\mathcal{C}|/\epsilon)). \tag{44}
$$

This finishes the proof. $\square$

5 Existence of The Small Cover

5.1 Intuition for Proof of Lemma 3

Given a space of degree-d polynomials $\mathcal{P}$, which is defined on $\mathbb{R}^m$ with co-dim at most k. Let $\mathcal{S}$ be defined as in Equation 33. Let $f(k, d, m, \epsilon)$ denote the size of the ϵ-cover for $\mathcal{S}$, which we aim to upper bound. The proof is finished via induction on $k + d + m$. For simplicity, here we omit the proof for the base case.

At every induction step with $k+d+m=I$, we assume the following holds for all $k'+d'+m' \le I-1$ and any ϵ

$$f(k', d', m', \epsilon) \le (2Rd'k'm'/\epsilon)^{C^2(d')^2(k')^{1/(d')}}. \tag{45}$$

And we prove the following for any (k,d,m) s.t. $k+d+m=I$ and any ϵ

$$f(k, d, m, \epsilon) \le (2Rdkm/\epsilon)^{C^2 d^2 k^{1/d}}. \tag{46}$$

Now for any (k,d,m), we want to cover the points $x \in \mathbb{R}^m$ where all polynomials $p \in \mathcal{P}$ nearly vanish. We first decompose $\mathbb{R}^m$ into $\mathbb{R}^{m'} \times \mathbb{R}^{m-m'}$, where $\mathbb{R}^{m'}$ is the space of the first m' coordinates. We will specify the value for m' later. Instead of considering all polynomials in $\mathcal{P}$, we only consider the polynomials in $\mathcal{P}$ that are degree-1 in the first m' coordinates, denoted by $\mathcal{P}_{m-m'}$. Namely, we cover a point as long as a smaller set of polynomials, i.e. $\mathcal{P}_{m-m'} \subset \mathcal{P}$, vanishes on it. This only enlarges our cover without losing anything. Notice here the co-dimension of $\mathcal{P}_{m-m'}$ in $\mathbb{R}^{m-m'}_{[d-1]}$ is still upper bounded by k, otherwise the co-dimension of $\mathcal{P}$ would have been greater than k, which contradicts with the definition of $\mathcal{P}$.

To construct the cover, we first brute-forcely covering space $\mathbb{R}^{m'}$. For $\forall x$ in our cover of $\mathbb{R}^{m'}$, we consider the corresponding polynomial space $\mathcal{P}_{m-m'}(x)$, where every polynomial is calculated by taking the corresponding polynomial in $\mathcal{P}_{m-m'}$ and evaluating its first m' coordinates on x. Now we need to find points in cylinder $\{x\} \times \mathbb{R}^{m-m'}$ where all polynomials in $\mathcal{P}_{m-m'}(x)$ vanishes, and cover them. The size of points in this cylinder is then $f(k, d-1, m-m', \epsilon)$ by definition. This "simple one-step recursion" establishes many of the intuitions we needed for the induction. However, there are several points that need to be discussed.

5.1.1 More on The "Simple One-Step Recursion"

At this point, one may wonder:

What if all polynomials in $\mathcal{P}_{m-m'}$ vanishes when evaluating its first m' coordinates on x?

When this happens, indeed, we need to cover the entire cylinder $\{x\} \times \mathbb{R}^{m-m'}$, which gives cover size exponential in $m-m' \approx m$, and fails our attempt to find small covers. *But in fact, this will never happen.* If the entire space $\mathcal{P}_{m-m'}$ vanishes on x, then the space of degree-$(d-1)$ polynomials defined on the variety of $\mathcal{P}_{m-m'}$ will have the same dimension as $\mathbb{R}^{m-m'}_{[d-1]}$, which is much larger than k itself. This contradicts with the fact that co-dimension of $\mathcal{P}_{m-m'}$ is upper bounded by k.

Another important point to note is that:

Simply repeating the above recursion is not enough!

To see this, we can directly write out the one-step recursion

$$f(k, d, m, \epsilon) = \left(\frac{R}{\epsilon}\right)^{m'} f(k, d-1, m-m', \epsilon). \tag{47}$$

Here the first factor comes from brute-forcely cover the space of the first m' coordinates. Applying this inequality for $d-1$ steps gives

$$f(k, d, m, \epsilon) = \left(\frac{R}{\epsilon}\right)^{m'(d-1)} f(k, 1, m-m', \epsilon). \tag{48}$$

Now applying the induction hypothesis on $f(k, 1, m-m', \epsilon)$ gives

$$f(k, d, m, \epsilon) = \left(\frac{R}{\epsilon}\right)^{m'(d-1)} \left(\frac{2Rdkm}{\epsilon}\right)^{C^2 k}. \tag{49}$$

This gives an upper bound of $f(k, d, m, \epsilon)$ which is still exponential in k. This implies that this simple "one-step recursion" ruins our efforts. We need to be more careful when considering the new $(d-1)$-degree polynomial space and its ϵ-cover.

5.1.2 A More Sophisticated Recursion

When covering $\mathbb{R}^{m'}$, we split all x's in this cover into two parts. For the **good points**, we require the co-dimension of the new polynomial space $\mathcal{P}_{m-m'}(x)$ to decrease to at most $k' = k^{1-1/d}$. Intuitively, since we are reducing the number of variables, less freedom is granted to the polynomials, and therefore dimension of the variety space we are covering should also be reduced. Otherwise, the point x doesn't tell us much about the variety space. In other words, the "information gain" is small from this specific point. We call such points where co-dim $\left(\mathcal{P}_{m-m'}(x)\right) > k^{1-1/d}$ **bad points**.

Clustering the good and bad points gives

$$\begin{aligned} f(k, d, m, \epsilon) = &\left(\frac{R}{\epsilon}\right)^{m'} f(k^{1-1/d}, d-1, m-m', \epsilon) \\ &+ \# \left(\text{cover for the cylinders of bad points}\right). \end{aligned} \tag{50}$$

If we ignore the second term, after $d-1$ steps, the first term decreases to

$$\begin{aligned}
&\left(\frac{R}{\epsilon}\right)^{m'(d-1)} f(k^{1/d}, 1, m-dm', \epsilon) \\
\overset{(i)}{\leq}\ &\left(\frac{R}{\epsilon}\right)^{m'(d-1)} \left(\frac{Rkdm}{\epsilon}\right)^{k^{1/d}} \\
\leq\ &\frac{1}{2m}\left(\frac{Rkdm}{\epsilon}\right)^{d\max\{k^{1/d}, m'\}},
\end{aligned} \tag{51}$$

which is no longer a problem. Here (i) is from the induction hypothesis.

On the other hand, we can show that the bad points actually concentrate together, which makes the second term small. Then combining the two terms finishes the induction and therefore finishes the proof of this lemma.

5.2 Proof of Lemma 3

Now we upper bound the second term in Equation 50, i.e. the cover size needed for the cylinders associated with the "bad points". This, combined with Equation 50, almost finishes the induction. To do this, we first show that bad points concentrate around a $2\lceil k^{1/d}\rceil$-dimensional hyperplane

Lemma 5 (A rough version of Lemma 22 in [DK20])**.** There exists a subspace $H \in \mathbb{R}^{m'}$ with dimension at most $2\lceil k^{1/d}\rceil$ so that all the bad points are close to H.

5.2.1 Rough Intuition for Lemma 5

The proof is finished by contradiction. We now talk about the intuition behind it. If there doesn't exist such hyperplane H, there must exist many bad x_i's such that x_i is far from the span of $\{x_1, \dots, x_{i-1}\}$. Here $i \in [\lceil 2k/k' \rceil]$.

For every x_i, we can at least find $k' = k^{1-1/d}$ different polynomials orthogonal to polynomials in $\mathcal{P}_{m-m'}(x_i)$. This is due to the definition of "bad points". One important fact is that:

From these polynomials, we can construct a $k' \cdot \lceil 2k/k' \rceil$-dimensional polynomial space that is almost orthogonal to polynomials in $\mathcal{P}_{m-m'}$.

Denote the above polynomial space as $\mathcal{Q}$. On the other hand, notice that this polynomial space has dimension lower bounded by $2k/k' * k' = 2k > k$. Since the co-dimension of $\mathcal{P}_{m-m'}$ is small (at most k), there must exist some polynomial $p \in \mathcal{P}_{m-m'}$ that also lies Q. Since all polynomials in Q are almost orthogonal to polynomials in $\mathcal{P}_{m-m'}$, this specific p must be almost orthogonal to itself, which causes a contradiction.

5.2.2 A Rough Proof of Lemma 5

Let t denote $2\lceil k^{1/d}\rceil$. If there doesn't exist such hyperplane H, as a simplifying assumption, we assume that we can find orthonormal vectors $\{x_1, \ldots, x_t\}$ such that the variety of $\mathcal{P}_{m-m'}(x_i)$ has co-dimension at least $k' = k^{1-1/d}$ for all $i \in [t]$ [2]. Namely, for every such x_i, we can find orthogonal degree $d-1$ polynomials $\{p_{i,1}, \ldots, p_{i,k'}\}$ such that

$$\left\|p_{i,j}\right\| = 1, \quad \langle p_{i,j}, p\rangle = 0, \forall p \in \mathcal{P}_{m-m'}(x_i). \tag{52}$$

With $\{p_{i,j}\}$, we now define the following polynomials $\{B_{i,j}\}$

$$B_{i,j}(x,y) = (x_i^\top x)p_{i,j}(y) \coloneqq q_i(x)p_{i,j}(y), \quad x \in \mathbb{R}^{m'}, y \in \mathbb{R}^{m-m'}. \tag{53}$$

It is then clear (from Claim 21 in [DK20]) that

$$\begin{aligned} \langle B_{i_1,j_1}, B_{i_2,j_2}\rangle &= \frac{1}{d}\langle q_{i_1}, q_{i_2}\rangle\langle p_{i_1,j_2}, p_{i_2,j_2}\rangle = 0, \quad \forall (i_1,j_1) \neq (i_2,j_2), \\ \langle B_{i,j}, B_{i,j}\rangle &= \frac{1}{d}\langle q_i, q_i\rangle\langle p_{i,j}, p_{i,j}\rangle \geq \frac{1}{d}, \quad \forall (i,j). \end{aligned} \tag{54}$$

For any polynomial $p \in \mathcal{P}_{m-m'}$ and any $i \in [t]$, we decompose it as $p = q_i(x)p_y(y) + \sum_{\tilde{p}_x} \tilde{p}_x(x)\tilde{p}_y(y)$ where $\tilde{p}_x$ are degree-1 polynomials on $\mathbb{R}^{m'}$ orthogonal to q_i. Therefore, we know that $\langle p, B_{i,j}\rangle = \frac{1}{d}\langle q_i, q_i\rangle\langle p_y, p_{i,j}\rangle$. On the other hand, p_y can be viewed as the polynomial got from p by evaluating the first m' coordinates on point x_i. This gives $\langle p_y, p_{i,j}\rangle = 0$, i.e.

$$\langle p, B_{i,j}\rangle = 0, \quad \forall p \in \mathcal{P}_{m-m'}. \tag{55}$$

On the other hand, notice that space $\mathcal{P}_{m-m'}$ has codimension at most k. However, the polynomial space spanned by $\{B_{i,j}\}$ has dimension $t \cdot k' \geq 2k > k$. Therefore, there must exist some polynomial $p \in \mathcal{P}_{m-m'}$ that can be written as $p = \sum_{i,j} \alpha_{i,j}B_{i,j}$. For this specific polynomial p, we know that there exist some (i,j) such that

$$\langle p, B_{i,j}\rangle \geq \alpha_{i,j} > 0. \tag{56}$$

This inequality contradicts with Equation 55. Therefore, there must exist such hyperplane H with dimension at most t.

[2] We refer the readers to the original paper for a more rigorous proof. However, following this simplifying assumption, the proof captures most of the essence of the original one.

5.2.3 Finishing The Recursion

With the above lemma, the space spanned by all the bad points and their cylinders has dimension $m - m' + 2\lceil k^{1/d}\rceil$. We need to cover points in this new space where the set of degree-d polynomial vanishes. Note that here we don't brute-forcely cover the bad points. Instead, we cluster them and their cylinders together to form a new space with lower dimension. And therefore the degree of the polynomial space doesn't decrease.

Now we are in the position to complete the induction. From the above lemma, we know that

$$\#\left(\text{cover for the cylinders of bad points}\right) = f(k, d, m - m' + \lceil 2k^{1/d}\rceil, \epsilon). \tag{57}$$

We choose $m' = \lceil 3k^{1/d}\rceil$ so that $m - m' + \lceil 2k^{1/d}\rceil \leq m - 1$. Therefore,

$$\#\left(\text{cover for the cylinders of bad points}\right) \leq f(k, d, m - 1, \epsilon). \tag{58}$$

Plugging back into Equation 50 gives

$$\begin{aligned}
f(k, d, m, \epsilon) &= \left(\frac{R}{\epsilon}\right)^{m'} f(k^{1-1/d}, d-1, m - m', \epsilon) + f(k, d, m-1, \epsilon) \\
&\overset{(i)}{\leq} \left(\frac{R}{\epsilon}\right)^{m'} \left(\frac{2Rdk^{1-1/d}m}{\epsilon}\right)^{C^2(d-1)^2(k^{1-1/d})^{\frac{1}{d-1}}} + \left(\frac{2Rdk(m-1)}{\epsilon}\right)^{C^2(d-1)^2k^{1/d}} \\
&\leq \left(\frac{R}{\epsilon}\right)^{3k^{1/d}} \left(\frac{2Rdk^{1-1/d}m}{\epsilon}\right)^{C^2(d-1)^2k^{1/d}} + \left((1-\frac{1}{m})\frac{2Rdkm}{\epsilon}\right)^{C^2d^2k^{1/d}} \\
&\leq \left(\frac{2Rdkm}{\epsilon}\right)^{C^2(d^2-d)k^{1/d}} + \left((1-\frac{1}{m})\frac{2Rdkm}{\epsilon}\right)^{C^2d^2k^{1/d}} \\
&\leq \frac{1}{2m}\left(\frac{2Rdkm}{\epsilon}\right)^{C^2d^2k^{1/d}} + (1-\frac{1}{m})\left(\frac{2Rdkm}{\epsilon}\right)^{C^2d^2k^{1/d}}
\end{aligned} \tag{59}$$

Finally, combining the terms gives

$$f(k, d, m, \epsilon) \leq \left(\frac{2Rdkm}{\epsilon}\right)^{C^2d^2k^{1/d}}, \tag{60}$$

which finishes the proof!

5.3 Algorithm to Construct The Small Cover

Algorithm 1: Construct The Small Cover of $\mathcal{P}$ on $\mathbb{R}^m$ with Paramters k, d, m

Init: k, d, m, $m' \leftarrow 3\lceil k^{1/d} \rceil$, $k' \leftarrow k^{1-1/d}$, Polynomial Space $\mathcal{P}$;

1 **if** $d = 1$ **then**

2   Directly solve for the "variety" of $\mathcal{P}$ on $\mathbb{R}^m$;

3 Construct a naive ϵ-cover for $\mathbb{R}^{m'}$ defined by the first m' coordinates, denoted by $\tilde{\mathcal{C}}$;

4 Cluster the good points and bad points in $\tilde{\mathcal{C}}$;

5 **for** *x is a good point* **do**

6     $\tilde{\mathcal{P}} \leftarrow P_{m-m'}(x)$;

7   Construct The Small Cover of $\tilde{\mathcal{P}}$ on $\mathbb{R}^{m-m'}$ with Parameters $k', d-1, m-m'$, denoted by $\tilde{C}_x$;

8 Construct the hyperplane H close to all bad points;

9 Construct The Small Cover of $\mathcal{P}$ on $H \times \mathbb{R}^{m-m'}$ with Parameters $k, d, m-1$, denoted by $\tilde{C}_{\text{bad}}$;

10 Return $\{\tilde{C}_x\}_{x \text{ is a good point}} \cup \tilde{C}_{\text{bad}}$;

From the above induction procedure, we automatically get the algorithm for constructing small covers for $\mathcal{S}$. Here line 2 is solvable because when $d = 1$, solving for the "variety" of $\mathcal{P}$ is equivalent to solving for a linear space (as discussed in Section 3.1). It is clear that the algorithm can run within time polynomial in the size of the cover, which is quasi-polynomial.

6 Conclusions & Future Directions

In this report, we start from a popular method in previous literature, which can be seen as calculating the variety for some degree-1 polynomial space. This type of method has an inherent exponential dependence on k. Using ideas from [DK20], we know that by extending the algorithm to calculate the variety for some higher-degree polynomial space, we can get an algorithm with quasi-polynomial sample and time complexity.

It is intriguing to look for polynomial algorithms in related fields based on the current method. One way to achieve this is to combine the small cover method with the techniques mentioned in [LL22]. Another approach is to go beyond the Method of Moments and try other techniques including Fourier moments [CLS20], etc.

7 Acknowledgments

I would like to express my sincere gratitude to Sitan for scheduling extra office hours to answer my questions and providing many intriguing research ideas in related fields. Also, I am grateful to my friend Tianqi Liu for patiently answering my elementary algebraic geometry questions.

References

[CLS20] Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–600, 2020.

[DK20] Ilias Diakonikolas and Daniel M Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 184–195. IEEE, 2020.

[DKS17] 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, 2017.

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018.

[DSS18] Ilias Diakonikolas, Anastasios Sidiropoulos, and Alistair Stewart. A polynomial time algorithm for maximum likelihood estimation of multivariate log-concave densities. *arXiv preprint arXiv:1812.05524*, 2018.

[HL18] 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, 2018.

[KSS18] 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, 2018.

[LL22] Allen Liu and Jerry Li. Clustering mixtures with almost optimal separation in polynomial time. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261, 2022.

[SOAJ14] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. *Advances in Neural Information Processing Systems*, 27, 2014.