Learning Mixtures of Gaussians with Censored Data

Wai Ming Tai [1] **Bryon Aragam** [1]

Abstract

We study the problem of learning mixtures of Gaussians with censored data. Statistical learning with censored data is a classical problem, with numerous practical applications, however, finite-sample guarantees for even simple latent variable models such as Gaussian mixtures are missing. Formally, we are given censored data from a mixture of univariate Gaussians

$$\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma^2),$$

i.e. the sample is observed only if it lies inside a set S. The goal is to learn the weights w_i and the means μ_i. We propose an algorithm that takes only $\frac{1}{\varepsilon^{O(k)}}$ samples to estimate the weights w_i and the means μ_i within ε error.

1. Introduction

When we collect data, we often encounter situations in which the data are partially observed. This can arise for a variety of reasons, such as measurements falling outside of the range of some apparatus or device. In machine learning and statistics, this phenomenon is known as truncated or censored data. Both refer to the case where we do not observe the data when they fall outside a certain domain. For censored data, we know the existence of data that fall outside the domain, while for truncated data, we do not.

It is common to encounter truncated or censored data in our daily lives. An example of truncated data is census data. When a census bureau collects data, there may be some difficulties for the bureau in collecting data for certain demographics for security, privacy, or legal reasons, and these individuals may have no incentive to report their data. Therefore, the bureau cannot collect data about these populations. In this case, the census data are truncated.

On the other hand, an example of censored data is test scores. The range of scores in a test is typically set to be from 0 to 100. Some students may score the maximum score of 100 points, in which case it is unknown if they could have scored even higher if the upper bound of the test score was higher than 100. Of course, even though the students' scores are capped at 100, their scores are still being reported. Hence, their scores are censored, which distinguishes them from truncated data.

Indeed, statistical estimation on truncated or censored data is a classical problem, dating back to the eighteenth century (Bernoulli, 1760). After Bernoulli, (Galton, 1898; Pearson, 1902; Pearson & Lee, 1908; Lee, 1914; Fisher, 1931) studied how to estimate the mean and the variance of of a univariate Gaussian distribution from truncated samples. However, most existing results do not address the problem of finite-sample bounds, i.e. the results are mostly experimental or asymptotic (Lee & Scott, 2012; McLachlan & Jones, 1988). In fact, one can learn the distribution with infinitely many truncated or censored samples—under mild assumptions, one can show that the function restricted on a certain region can be extended to the entire space by the identity theorem from complex analysis. Unfortunately, it is still not clear how to translate such results to finite sample bounds.

A recent notable result by Daskalakis et al. (2018) gave the first efficient algorithm to learn the mean and the covariance of a single Gaussian with finitely many truncated samples. A natural extension to the problem of learning a single Gaussian is the problem of learning a mixture of Gaussians. To the best of our knowledge, there is no provable guarantees on the problem of learning a mixture of Gassians with a finite number of truncated or censored samples even in one dimension.

As we will discuss in the related work section, there is a long line of work on learning a mixture of Gaussians. Likelihood-based approaches often do not provide provable guarantees for learning mixtures of Gaussians since the objective function is not convex unless we impose strong assumptions (Xu et al., 2016; Daskalakis et al., 2017). On the other hand, many recent results rely heavily on the method of moments, i.e. the algorithm estimates the moments $\mathsf{E}(X^s)$

[1] Booth School of Business, University of Chicago, Chicago, USA. Correspondence to: Wai Ming Tai <waiming.tai@chicagobooth.edu>, Bryon Aragam <bryon@chicagobooth.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

as an intermediate step. With truncated or censored data, estimating $\mathsf{E}(X^s)$ (here, the expectation is over the original, untruncated data) becomes very challenging.

To overcome this, we propose an approach for estimating moments from censored data. Recall that ordinary moments are just expectations of monomials of a random variable. However, by generalizing this to more general functions of a random variable, we open up the possibility to capture more complex structures of the distribution. In particular, when the data is censored, these generalized functions allow us to relate the expectations back the raw, uncensored distribution. One must keep in mind that we still need to make a choice of what functions to consider in addition to providing efficient estimators of these generalized moments. We preview that a suitable choice is found by a specific linear combination of Hermite polynomials derived from the solution to a system of linear equations. In our proof, we will delve deeper into the analysis of the expectations of functions, depending on the domain, and provide a delicate analysis to prove our desired result.

Based on the above discussion, we may want to ask the following question in a general sense: *Can we learn a mixture of Gaussians with truncated or censored data?* In this paper, we consider this problem and focus on the case that the data is censored and the Gaussians are univariate and homogeneous. We now define the problem formally.

2. Problem Definition

Let $\mathcal{N}(\mu, \sigma^2)$ be the normal distribution with mean μ and variance σ^2. Namely, the pdf of $\mathcal{N}(\mu, \sigma^2)$ is

$$g_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

For any subset $S \subset \mathbb{R}$, let $I_{\mu,\sigma^2}(S)$ be the probability mass of $\mathcal{N}(\mu, \sigma^2)$ on S, i.e.

$$I_{\mu,\sigma^2}(S) := \int_{x \in S} g_{\mu,\sigma^2}(x)\mathsf{d}x.$$

Also, let $\mathcal{N}(\mu, \sigma^2, S)$ denote the conditional distribution of a normal $\mathcal{N}(\mu, \sigma^2)$ given the set S. Namely, the pdf of $\mathcal{N}(\mu, \sigma^2, S)$ is

$$g_{\mu,\sigma^2,S}(x) := \begin{cases} \frac{1}{I_{\mu,\sigma^2}(S)} g_{\mu,\sigma^2}(x) & \text{if } x \in S \\ 0 & \text{if } x \notin S. \end{cases}$$

Given a subset $S \subset \mathbb{R}$, we consider the following sampling procedure. Each time, a sample is drawn from a mixture of Gaussians

$$\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma^2),$$

where $w_i > 0$, $\sum_{i=1}^k w_i = 1$, $\mu_i \in \mathbb{R}$, and $\sigma > 0$. If this sample is inside S, we obtain this sample; otherwise, we fail to generate a sample. Formally, X is a random variable drawn from the following distribution. Let α be the probability mass $\sum_{i=1}^k w_i I_{\mu_i,\sigma^2}(S)$.

$$X \sim \begin{cases} \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \sigma^2, S) & \text{with probability } \alpha \\ \mathsf{FAIL} & \text{with probability } 1 - \alpha. \end{cases} \tag{1}$$

The value FAIL here refers to values that are not directly accessible to the algorithm.

We assume that

(A1) S is an interval $[-R, R]$ for some constant $R > 0$ and is known (it is easy to extend S to be any measurable subset of $[-R, R]$; for simplicity, we assume $S = [-R, R]$);

(A2) All μ_i are bounded, i.e. $|\mu_i| < M$ for some constant $M > 0$;

(A3) The variance σ^2 is known.

We also assume that the exact computation of the integral $\int_0^z e^{-\frac{1}{2}t^2}\mathsf{d}t$ for any z can be done. Indeed, one can always approximate this integral with an exponential convergence rate via Taylor expansion. As we can see in our proof, this error is negligible.

For a given error parameter $\varepsilon > 0$, we want to estimate all w_i, μ_i within ε error. The question is *how many samples from the above sampling procedure do we need to achieve this goal?* Our main contribution is a quantitative answer to this question. We will prove the following theorem:

Theorem 2.1. *Suppose we have n samples drawn from the distribution (1) and we assume that the mixture satisfies (A1)-(A3). Furthermore, let $w_{\min}$ be $\min\{w_i \mid i = 1, \ldots, k\}$ and $\Delta_{\min}$ be $\min\{|\mu_i - \mu_j| \mid i, j = 1, \ldots, k \text{ and } i \neq j\}$. Then, for a sufficiently small $\varepsilon > 0$, if $w_{\min}$ and $\Delta_{\min}$ satisfy $w_{\min}\Delta_{\min} = \Omega(\varepsilon)$, there is an efficient algorithm that takes $n = C_k \cdot \frac{1}{\varepsilon^{O(k)}}$ (where C_k is a constant depending on k only) samples[1] as the input and outputs $\widehat{w}_i, \widehat{\mu}_i$ for $i = 1, \ldots, k$ such that, up to an index permutation Π,*

$$|\widehat{w}_{\Pi(i)} - w_i| < \varepsilon, \qquad |\widehat{\mu}_{\Pi(i)} - \mu_i| < \varepsilon \qquad \text{for } i = 1, \ldots, k$$

with probability $\frac{99}{100}$. The running time of the algorithm is $O(n \cdot \mathit{poly}(k, \frac{1}{\varepsilon}))$.

[1] Here we assume the parameters R and M to be constant for simplicity. It is easy to keep track of them in our proof and show that the sample bound is $C_k \cdot (\frac{1}{\varepsilon})^{O(k \cdot \log(M + R + \frac{1}{R}))}$.

In other words, this theorem states that the sample complexity for learning mixtures of k univariate Gaussians with censored data is $\frac{1}{\varepsilon^{O(k)}}$ which is optimal in terms of asymptotic growth of the exponent $O(k)$ (Wu & Yang, 2018). As for the optimality of the constant in the exponent $O(k)$, this is an interesting open problem.

3. Related Work

Without truncation or censoring, the study of learning Gaussian mixture models (Pearson, 1894) has a long history. We focus on recent algorithmic results; see Lindsay (1995) for additional background. Dasgupta (1999) proposed an algorithm to learn the centers of each Gaussian when the centers are $\Omega(\sqrt{d})$ apart from each other. There are other results such as (Vempala & Wang, 2004; Regev & Vijayaraghavan, 2017) that are based on similar separation assumptions and that use clustering techniques.

There are other results using the method of moments. Namely, the algorithm estimates the moments $\mathsf{E}(X^s)$ as an intermediate step. Moitra & Valiant (2010); Kalai et al. (2010) showed that, assuming $k = O(1)$, there is an efficient algorithm that learns the parameters with $\frac{1}{\varepsilon^{O(k)}}$ samples. Hardt & Price (2015) showed that, when $k = 2$, the optimal sample complexity of learning the parameters is $\Theta(\frac{1}{\varepsilon^{12}})$. For the case that the Gaussians in the mixture have equal variance, Wu & Yang (2018) proved the optimal sample complexity for learning the centers is $\Theta(\frac{1}{\varepsilon^{4k-2}})$ if the variance is known and $\Theta(\frac{1}{\varepsilon^{4k}})$ if the variance is unknown. Later, Doss et al. (2020) extended the optimal sample complexity to high dimensions.

When the data are truncated or censored, however, the task becomes more challenging. (Schneider, 1986; Balakrishnan & Cramer, 2014; Cohen, 2016) provided a detailed survey on the topic of learning Gaussians with truncated or censored data. Recently, Daskalakis et al. (2018) showed that, if the samples are from a single Gaussian in high dimensional spaces, there is an algorithm that uses $\widetilde{O}(\frac{d^2}{\varepsilon^2})$ samples to learn the mean vector and the covariance matrix. Their approach is likelihood based. Namely, they optimize the negative log-likelihood function to find the optimal value. This approach relies on the fact that, for a single Gaussian, the negative log-likelihood function is convex and hence one can use greedy approaches such as stochastic gradient descent to find the optimal value.

Unfortunately, when there are multiple Gaussians in the mixture, we may not have such convexity property for the negative log-likelihood function. Nagarajan & Panageas (2020) showed that, for the special case of a truncated mixture of two Gaussians whose centers are symmetric around the origin and assuming the truncated density is known, the output by the EM algorithm converges to the true mean as the number of iterations tends to infinity.

There are other problem settings that are closely related to ours such as robust estimation of the parameters of a Gaussian in high dimensional spaces. The setting of robust estimation is the following. The samples we observed are generated from a single high dimensional Gaussians except that a fraction of them is corrupted. Multiple previous results such as (Hopkins et al., 2022; Liu et al., 2021; Diakonikolas & Kane, 2019; Diakonikolas et al., 2019; Lai et al., 2016; Diakonikolas et al., 2017; 2018) proposed learning algorithms to learn the mean vector and the covariance matrix.

Regression with truncated or censored data is another common formulation. Namely, we only observe the data when the value of the dependent variable lies in a certain subset. A classic formulation is the truncated linear regression model (Tobin, 1958; Amemiya, 1973; Hausman & Wise, 1977; Maddala, 1986). Recently, in the truncated linear regression model, Daskalakis et al. (2019) proposed a likelihood-based estimator to learn the parameters.

4. Preliminaries

We denote the set $\{0, 1, \ldots, n-1\}$ to be $[n]$ for any positive integer n. Let $h_j(x)$ be the (probabilist's) Hermite polynomials, i.e.

$$h_j(x) = (-1)^j e^{\frac{1}{2}x^2} \frac{\mathsf{d}^j}{\mathsf{d}\xi^j} e^{-\frac{1}{2}\xi^2}\Big|_{\xi=x} \qquad \text{for all } x \in \mathbb{R}.$$

Hermite polynomials can also be given by the exponential generating function, i.e.

$$e^{x\mu - \frac{1}{2}\mu^2} = \sum_{j=0}^{\infty} h_j(x) \frac{\mu^j}{j!} \qquad \text{for any } x, \mu \in \mathbb{R}. \qquad (2)$$

Also, the explicit formula for h_j is

$$h_j(x) = j! \sum_{i=0}^{\lfloor j/2 \rfloor} \frac{(-1/2)^i}{i!(j-2i)!} x^{j-2i}$$

and this explicit formula is useful in our analysis.

In our proof, we will solve multiple systems of linear equations. Cramer's rule provides an explicit formula for the solution of a system of linear equations whenever the system has a unique solution.

Lemma 4.1 (Cramer's rule)**.** *Consider the following system of n linear equations with n variables.*

$$Ax = b$$

where A is a n-by-n matrix with nonzero determinant and b is a n dimensional vector. Then, the solution of this system

$\widehat{x} = A^{-1}b$ satisfies that the i-th entry of $\widehat{x}$ is

$$\det(A^{(i \leftarrow b)}) / \det(A)$$

where $A^{(i \leftarrow b)}$ is the same matrix as A except that the i-th column is replaced with b.

Thanks to the application of Cramer's rule, we often encounter determinants. The Cauchy-Binet formula is a formula for the determinant of a matrix that each entry can be expressed as an inner product of two vectors that correspond to its row and column. Note that the Cauchy-Binet formula usually applies to the case that the entries are finite sums. For our purpose, we state the Cauchy-Binet formula for the case that the entries are in integral form.

Lemma 4.2 (Cauchy–Binet formula)**.** *Let A be a n-by-n matrix whose (r, c)-entry has a form of $\int_{x \in S} f_r(x) g_c(x) \mathrm{d}x$ for some functions f_r, g_c and some domain $S \subset \mathbb{R}$. Then, the determinant of A is*

$$\det(A) = \int_{x_0 > \cdots > x_{n-1}, \mathbf{x} \in S^n} \det(B(\mathbf{x})) \cdot \det(C(\mathbf{x})) \mathrm{d}\mathbf{x}$$

where, for any $\mathbf{x} = (x_0, \ldots, x_{n-1}) \in S^n$, $B(\mathbf{x})$ is a n-by-n matrix whose (r, i)-entry is $f_r(x_i)$ and $C(\mathbf{x})$ is a n-by-n matrix whose (i, c)-entry is $g_c(x_i)$.

Another tool to help us compute the determinants is Schur polynomials. Schur polynomials are defined as follows. For any partition $\lambda = (\lambda_1, \ldots, \lambda_n)$ such that $\lambda_1 \geq \cdots \geq \lambda_n$ and $\lambda_i \geq 0$, define the function $a_{(\lambda_1+n-1, \lambda_2+n-2, \ldots, \lambda_n)}(x_1, x_2, \ldots, x_n)$ to be

$$a_{(\lambda_1+n-1, \lambda_2+n-2, \ldots, \lambda_n)}(x_1, x_2, \ldots, x_n) := \det \begin{bmatrix} x_1^{\lambda_1+n-1} & x_2^{\lambda_1+n-1} & \cdots & x_n^{\lambda_1+n-1} \\ x_1^{\lambda_2+n-2} & x_2^{\lambda_2+n-2} & \cdots & x_n^{\lambda_2+n-2} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{\lambda_n} & x_2^{\lambda_n} & \cdots & x_n^{\lambda_n} \end{bmatrix}.$$

In particular, when $\lambda = (0, 0, \ldots, 0)$, it becomes the Vandermonde determinant, i.e.

$$a_{(n-1, n-2, \ldots, 0)}(x_1, x_2, \ldots, x_n) = \prod_{1 \leq j < k \leq n} (x_j - x_k).$$

Then, Schur polynomials are defined to be

$$s_\lambda(x_1, x_2, \ldots, x_n) := \frac{a_{(\lambda_1+n-1, \lambda_2+n-2, \ldots, \lambda_n)}(x_1, x_2, \ldots, x_n)}{a_{(n-1, n-2, \ldots, 0)}(x_1, x_2, \ldots, x_n)}.$$

It is known that $s_\lambda(x_1, x_2, \ldots, x_n)$ can be written as $\sum_Y \mathbf{x}^Y$ where the summation is over all semi-standard Young tableaux Y of shape λ. Here, each term $\mathbf{x}^Y$ means $x_1^{y_1} \cdots x_n^{y_n}$ where y_i is the number of occurrences of the number i in Y and note that $\sum_{i=1}^n y_i = \sum_{i=1}^n \lambda_i$. Also, a semi-standard Young tableau Y of shape $\lambda = (\lambda_1, \ldots, \lambda_n)$ can be represented by a finite collection of boxes arranged in left-justified rows where the row length is λ_i and each box is filled with a number from 1 to n such that the numbers in each row is non-decreasing and the numbers in each column is increasing. To avoid overcomplicating our argument, when we count the number of semi-standard Young tableaux of some shape we only use a loose bound for it.

5. Proof Overview

Recall that our setting is the following (cf. (1)): We are given samples drawn from the following sampling procedure. Each time, a sample is drawn from a mixture of Gaussians

$$\sum_{i=1}^k w_i \mathcal{N}(\mu_i, \sigma^2)$$

where $w_i > 0, \sum_{i=1}^k w_i = 1, \mu_i \in \mathbb{R}$ and $\sigma > 0$. If this sample is inside S, we obtain this sample; otherwise, we fail to generate a sample. Our goal is to learn w_i and μ_i.

One useful way to view mixtures of Gaussians is to express it as

$$\left(\sum_{i=1}^k w_i \delta_{\mu_i} \right) * \mathcal{N}(0, \sigma^2)$$

where δ_{μ_i} is the delta distribution at μ_i and $*$ is the convolution operator. We call the distribution $\sum_{i=1}^k w_i \delta_{\mu_i}$ the mixing distribution. Let $\mathbf{m}_j$ be the moment of the mixing distribution, i.e.

$$\mathbf{m}_j := \sum_{i=1}^k w_i \mu_i^j.$$

Since we assume that the variance is known, without loss of generality, we set $\sigma = 1$; otherwise, we can scale all samples such that $\sigma = 1$. First, we reduce the problem to estimating $\mathbf{m}_j$, so that we can employ known results on estimating mixtures of Gaussians using the method of moments. For example, Wu & Yang (2018) proved the following theorem.

Theorem 5.1 (Denoised method of moments, (Wu & Yang, 2018))**.** *Suppose $\mathbf{m}_j$ are the moments of a distribution that has k supports on $\mathbb{R}$, i.e. $\mathbf{m}_j$ has a form of $\sum_{i=1}^k w_i \mu_i^j$ where $w_i > 0$, $\sum_{i=1}^k w_i = 1$ and $\mu_i \in \mathbb{R}$. Let $w_{\min}$ be $\min\{w_i \mid i = 1, \ldots, k\}$ and $\Delta_{\min}$ and $\min\{|\mu_i - \mu_j| \mid i, j = 1, \ldots, k \text{ and } i \neq j\}$. For any $\delta > 0$, let $\widehat{\mathbf{m}}_j$ be the numbers that satisfy*

$$|\widehat{\mathbf{m}}_j - \mathbf{m}_j| < \delta \qquad \text{for all } j = 1, \ldots, 2k-1.$$

Then, if $w_{\min}$ and $\Delta_{\min}$ satisfy $w_{\min}\Delta_{\min} = \Omega(\delta^{O(\frac{1}{k})})$, there is an algorithm that takes $\widehat{\mathbf{m}}_j$ as the input and outputs $\widehat{w}_i, \widehat{\mu}_i$ such that, up to an index permutation Π,

$$|\widehat{w}_{\Pi(i)} - w_i| < C_k \cdot \frac{\delta^{\Omega(\frac{1}{k})}}{w_{\min}}$$

and

$$|\widehat{\mu}_{\Pi(i)} - \mu_i| < C_k \cdot \frac{\delta^{\Omega(\frac{1}{k})}}{\Delta_{\min}}$$

where C_k is a constant depending on k only.

Unfortunately, unlike with fully observed mixtures of Gaussians, estimating these moments is no longer straightforward. As we will see in our proof, looking for unbiased estimators relies on specific structures of Gaussians. When the data is censored, such structures may not exist. Hence, we look for a biased estimator and provide delicate analysis to bound the bias. To see how we can estimate $\mathbf{m}_j$, we first express the mixture as an expression that is in terms of $\mathbf{m}_j$. Suppose X is the random variable drawn from the sampling procedure conditioned on non-FAIL samples. For any function f, the expectation of $f(X)$ is

$$\begin{aligned}\mathsf{E}(f(X)) &= \int_S f(x) \cdot \left(\sum_{i=1}^{k} w_i g_{\mu_i,1,S}(x)\right) \mathsf{d}x \\ &= \frac{1}{\alpha} \cdot \int_S f(x) \cdot \left(\sum_{i=1}^{k} w_i g_{\mu_i,1}(x)\right) \mathsf{d}x. \quad (3)\end{aligned}$$

Recall that α is the probability mass $\sum_{i=1}^{k} w_i I_{\mu_i,1}(S)$. Note that, for any μ,

$$\begin{aligned}g_{\mu,1}(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} e^{x\mu - \frac{1}{2}\mu^2} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \sum_{j=0}^{\infty} h_j(x) \frac{\mu^j}{j!} \quad (4)\end{aligned}$$

where h_j is the j-th Hermite polynomial and the last equality is from the fact (2). In other words, when we plug (4) into (3), we have

$$\begin{aligned}&\alpha \cdot \mathsf{E}(f(X)) \\ &= \int_S f(x) \cdot \left(\sum_{i=1}^{k} w_i \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \sum_{j=0}^{\infty} h_j(x) \frac{\mu_i^j}{j!}\right)\right) \mathsf{d}x \\ &= \sum_{j=0}^{\infty} \left(\int_S f(x) \cdot \frac{1}{\sqrt{2\pi} j!} e^{-\frac{1}{2}x^2} h_j(x) \mathsf{d}x\right) \cdot \left(\sum_{i=1}^{k} w_i \mu_i^j\right).\end{aligned}$$

To ease the notation, for any function f and positive integer j, we define

$$J_{f,j} := \int_S f(x) \cdot \frac{1}{\sqrt{2\pi} j!} e^{-\frac{1}{2}x^2} h_j(x) \mathsf{d}x. \quad (5)$$

If we plug $J_{f,j}$ and $\mathbf{m}_j$ into the equation for $\alpha \cdot \mathsf{E}(f(X))$, we have

$$\alpha \cdot \mathsf{E}(f(X)) = \sum_{j=0}^{\infty} J_{f,j} \cdot \mathbf{m}_j. \quad (6)$$

Ideally, *if* we manage to find $2k-1$ functions $f_1, \dots, f_{2k-1}$ such that

$$J_{f_i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

then we have

$$\alpha \cdot \mathsf{E}(f_i(X)) = \mathbf{m}_i \qquad \text{for all } i = 1, \dots, 2k-1.$$

It means that we will have an unbiased estimator for $\mathbf{m}_i$ and therefore we just need to find out the amount of samples we need by bounding the variance. Indeed, if $S = \mathbb{R}$ and we pick f_i to be the i-th Hermite polynomial h_i then the aforementioned conditions hold. It is how (Wu & Yang, 2018) managed to show their result. However, when $S \neq \mathbb{R}$, it becomes trickier.

A natural extension is to pick f_i to be a linear combination of Hermite polynomials, i.e.

$$f_i = \sum_{a=0}^{\ell-1} \beta_{i,a} h_a$$

for some positive integer ℓ. The integer ℓ is a parameter indicating how accurate our estimator is. Indeed, this $\ell \to \infty$ as $\varepsilon \to 0$ as we will show in our proof. For each f_i, there are ℓ coefficients $\beta_{i,j}$ in the expression and therefore we can enforce ℓ terms of $J_{f_i,j}$ to be the desired values. More precisely, we can set $\beta_{i,a}$ such that

$$\begin{aligned}J_{f_i,j} &= \int_S f_i(x) \cdot \frac{1}{\sqrt{2\pi} j!} e^{-\frac{1}{2}x^2} h_j(x) \mathsf{d}x \\ &= \sum_{a=0}^{\ell-1} \beta_{i,j} \int_S h_a(x) \cdot \frac{1}{\sqrt{2\pi} j!} e^{-\frac{1}{2}x^2} h_j(x) \mathsf{d}x \\ &= \sum_{a=0}^{\ell-1} \beta_{i,a} J_{h_a,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}\end{aligned}$$

for $j = 0, \dots, \ell-1$. If we assume the integrals can be computed exactly, then all $J_{h_a,j}$ are known. Hence, we can solve $\beta_{i,a}$ by solving this system of linear equations.

Now, if we plug them into (6) then we have

$$\alpha \cdot \mathsf{E}(f_i(X)) = \mathbf{m}_i + \underbrace{\sum_{j=\ell}^{\infty} J_{f_i,j} \cdot \mathbf{m}_j}_{:=\mathcal{E}_i} .$$

Note that the term $\mathbf{m}_i$ is what we aim at and hence the term $\mathcal{E}_i$ is the error term. Indeed, our estimator is a biased estimator where the bias is $\mathcal{E}_i$. Thanks to the factor $\frac{1}{j!}$ in the term $J_{f_i,j}$, intuitively, the term $\mathcal{E}_i \to 0$ as $\ell \to 0$.

Define our estimator $\widehat{\mathbf{m}}_i$ to be

$$\widehat{\mathbf{m}}_i = \frac{1}{n}\left(\sum_{s=1}^{n'} f_i(x_s)\right) \tag{7}$$

where n' is the number of samples that are non-FAIL and x_i are the non-FAIL samples. Note that, on average, the term $\frac{1}{n} = \frac{\alpha}{n'}$ gives us the factor α implicitly. Then, by Chebyshev's inequality, we have

$$|\widehat{\mathbf{m}}_i - \mathbf{m}_i| < \delta + |\mathcal{E}_i| \quad \text{with probability } 1 - \frac{\mathsf{Var}(\widehat{\mathbf{m}}_i)}{\delta^2}.$$

Now, we break the problem down to the following two subproblems.

- How large ℓ needs to be in order to make $|\mathcal{E}_i| < \delta$?
- Given $\delta > 0$, how many samples do we need to make the variance $\mathsf{Var}(\widehat{\mathbf{m}}_i) < \frac{\delta^2}{100}$ and hence the success probability larger than $\frac{99}{100}$?

Detailed proofs are deferred to the appendix.

5.1. Bounds for the Number of Terms

To see how large ℓ needs to be, we first define the following notations. Let $v^{(j)}$ be the ℓ-dimensional vector whose a-th entry is $J_{h_a,j}$, i.e.

$$v^{(j)} = \begin{bmatrix} J_{h_0,j} & J_{h_1,j} & \cdots & J_{h_{\ell-1},j} \end{bmatrix}^\top,$$

and V be the the ℓ-by-ℓ matrix whose r-th row is $(v^{(r)})^\top$, i.e.

$$V = \begin{bmatrix} v^{(0)} & v^{(1)} & \cdots & v^{(\ell-1)} \end{bmatrix}^\top. \tag{8}$$

Recall that, by the definition of $\beta_{i,a}$, $\beta_{i,a}$ satisfies

$$\sum_{a=0}^{\ell-1} \beta_{i,a} J_{h_a,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} .$$

We can rewrite it as a system of linear equations.

$$V\beta_i = \mathbf{e}_i \tag{9}$$

where β_i is the ℓ-dimensional vector whose a-th entry is $\beta_{i,a}$ and $\mathbf{e}_i$ is the ℓ-dimensional canonical vector which is a zero vector except that the i-th entry is 1, i.e.

$$\beta_i = \begin{bmatrix} \beta_{i,0} & \beta_{i,1} & \cdots & \beta_{i,\ell-1} \end{bmatrix}^\top$$

and

$$\mathbf{e}_i = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}^\top .$$

Namely, we have $\beta_i = V^{-1}\mathbf{e}_i$. Recall that the definition of $\mathcal{E}_i$ is

$$\mathcal{E}_i = \sum_{j=\ell}^{\infty} J_{f_i,j} \cdot \mathbf{m}_j .$$

To bound the term $J_{f_i,j}$, observe that

$$J_{f_i,j} = \sum_{a=0}^{\ell-1} \beta_{i,a} J_{h_a,j} = (v^{(j)})^\top V^{-1} \mathbf{e}_i$$

and, by Cramer's rule, $J_{f_i,j}$ can be expressed as

$$J_{f_i,j} = \frac{\det(V^{(i\to j)})}{\det(V)}$$

where $V^{(i\to j)}$ is the same matrix as V except that the i-th row is replaced with $v^{(j)}$, i.e.

$$\begin{aligned} &V^{(i\to j)} \\ &= \begin{bmatrix} v^{(0)} & \cdots & v^{(i-1)} & v^{(j)} & v^{(i+1)} & \cdots & v^{(\ell-1)} \end{bmatrix}^\top \end{aligned} \tag{10}$$

for $i = 1, \ldots, 2k-1$ and $j \geq \ell$. The right arrow in the superscript indicates the row replacement. We preview that there are column replacements in our calculation and we will use left arrows to indicate it.

In Lemma A.2, we show that

$$|J_{f_i,j}| = \frac{|\det(V^{(i\to j)})|}{|\det(V)|} \leq \frac{1}{2^{\Omega(j\log j)}}.$$

Also, by the assumption that $|\mu_i| < M$ where M is a constant, we have $\mathbf{m}_j \leq M^j$. Hence, we prove that

$$\begin{aligned} |\mathcal{E}_i| &\leq \sum_{j=\ell}^{\infty} |J_{f_i,j}||\mathbf{m}_j| \leq \sum_{j=\ell}^{\infty} \frac{1}{2^{\Omega(j\log j)}} \cdot M^j \\ &\leq \frac{1}{2^{\Omega(\ell\log\ell)}} \cdot M^\ell \leq \delta \end{aligned}$$

as long as $\ell = \Omega(\frac{\log\frac{1}{\delta}}{\log\log\frac{1}{\delta}})$.

Hence, we have the following lemma.

Lemma 5.2. *For a sufficiently small $\delta > 0$, when $\ell = \Omega(\frac{\log\frac{1}{\delta}}{\log\log\frac{1}{\delta}})$, the estimators $\widehat{\mathbf{m}}_i$ computed by (7) satisfies*

$$|\widehat{\mathbf{m}}_i - \mathbf{m}_i| < 2\delta \qquad \textit{with probability } 1 - \frac{\mathit{Var}(\widehat{\mathbf{m}}_i)}{\delta^2}.$$

5.2. Bounds for the Variance

Recall that our second subproblem is to bound the variance of our estimator. To bound $\mathsf{Var}(\widehat{\mathbf{m}}_i)$, observe that

$$\begin{aligned}\mathsf{Var}(\widehat{\mathbf{m}}_i) &\le \mathsf{E}(\widehat{\mathbf{m}}_i^2) = \frac{\alpha}{n}\mathsf{E}(f_i(X)^2)\\ &= \frac{\alpha}{n}\mathsf{E}\Big(\big(\sum_{a=0}^{\ell-1}\beta_{i,a}h_a(X)\big)^2\Big)\\ &\le \frac{\alpha}{n}\left(\sum_{a=0}^{\ell-1}|\beta_{i,a}|\sqrt{\mathsf{E}\left(h_a(X)^2\right)}\right)^2 \qquad (11)\end{aligned}$$

By expanding the expectation explicitly,

$$\begin{aligned}\mathsf{E}\left(h_a(X)^2\right) &= \int_S h_a(x)^2\cdot\left(\sum_{i=1}^{k} w_i g_{\mu_i,1,S}(x)\right)\mathsf{d}x\\ &\le \frac{1}{\alpha}\int_{\mathbb{R}} h_a(x)^2\cdot\left(\sum_{i=1}^{k} w_i g_{\mu_i,1}(x)\right)\mathsf{d}x\\ &\le \frac{1}{\alpha}(O(M+\sqrt{a}))^{2a} \qquad (12)\end{aligned}$$

The last line comes from (Wu & Yang, 2018) where they showed that

$$\int_{\mathbb{R}} h_a(x)^2\cdot\left(\sum_{i=1}^{k} w_i g_{\mu_i,1}(x)\right)\mathsf{d}x \le (O(M+\sqrt{a}))^{2a}$$

in Lemma 5 of (Wu & Yang, 2018).

Now, we also need to bound $|\beta_{i,a}|$. Recall that

$$\beta_i = V^{-1}\mathbf{e}_i.$$

By Cramer's rule, each coordinate of β_i is

$$\beta_{i,a} = \frac{\det(V^{(\mathbf{e}_i\leftarrow a)})}{\det(V)}$$

where $V^{(\mathbf{e}_i\leftarrow a)}$ is the same matrix as V except that the a-th column is replaced with $\mathbf{e}_i$, i.e.

$$\begin{aligned}&V^{(\mathbf{e}_i\leftarrow a)}\\ &= \begin{bmatrix} v_0^{(0)} & \cdots & v_{a-1}^{(0)} & 0 & v_{a+1}^{(0)} & \cdots & v_{\ell-1}^{(0)}\\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\ v_0^{(i)} & \cdots & v_{a-1}^{(i)} & 1 & v_{a+1}^{(i)} & \cdots & v_{\ell-1}^{(i)}\\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\ v_0^{(\ell-1)} & \cdots & v_{a-1}^{(\ell-1)} & 0 & v_{a+1}^{(\ell-1)} & \cdots & v_{\ell-1}^{(\ell-1)} \end{bmatrix}\\ &\qquad (13)\end{aligned}$$

In Lemma A.3, we show that

$$|\beta_{i,a}| \le 2^{O(\ell\log\ell)}. \qquad (14)$$

Therefore, if we plug (12) and (14) into (11), we have the following lemma.

Algorithm 1 Learning mixtures of Gaussians with censored data

Input: n iid samples $x_1,\ldots,x_n$, number of Gaussians k, parameter ℓ, mean boundary parameter M, sample domain $S=[-R,R]$

```
for $i = 0$ to $2k-1$ do
    solve (9) to obtain $\beta_i = (\beta_{i,0},\beta_{i,1},\ldots,\beta_{i,\ell-1})^\top$, i.e. solve the following system of linear equations
        $V\beta_i = \mathbf{e}_i$
    where the $(r,c)$-entry of $V$ is
        $\int_S \frac{1}{\sqrt{2\pi}r!}e^{-\frac{1}{2}x^2}h_c(x)h_r(x)\mathsf{d}x$    (15)
    and $\mathbf{e}_i$ is the canonical vector
end for
for each sample $x_s$ do
    compute $\widehat{f}_i(x_s) := \begin{cases} f_i(x_s) & \text{if } x_s \text{ is non-}\mathsf{FAIL}\\ 0 & \text{if } x_s \text{ is } \mathsf{FAIL}\end{cases}$;
    recall that $f_i$ is
        $f_i(x) = \sum_{a=0}^{\ell-1}\beta_{i,a}h_a(x)$
    and $h_a$ is the $a$-th Hermite polynomial
        $h_a(x) = a!\sum_{j=0}^{\lfloor a/2\rfloor}\frac{(-1/2)^j}{j!(a-2j)!}x^{a-2j}$
end for
for $i = 1$ to $2k-1$ do
    compute $\widehat{\mathbf{m}}_i = \frac{1}{n}\sum_{s=1}^{n}\widehat{f}_i(x_s)$ which is the same as the estimator defined in (7)
end for
let $\widehat{w}_1,\widehat{w}_2,\ldots,\widehat{w}_k$ and $\widehat{\mu}_1,\widehat{\mu}_2,\cdots,\widehat{\mu}_k$ be the output of Algorithm 2 using $\widehat{\mathbf{m}} = (\widehat{\mathbf{m}}_1,\ldots,\widehat{\mathbf{m}}_{2k-1})$ and $M$ as the input
```

Output: estimated weights $\widehat{w}_1,\widehat{w}_2,\ldots,\widehat{w}_k$ and estimated means $\widehat{\mu}_1,\widehat{\mu}_2,\cdots,\widehat{\mu}_k$

Lemma 5.3. *For any positive integer ℓ, the estimator $\widehat{\mathbf{m}}_i$ computed by (7) has variance*

$$\mathit{Var}(\widehat{\mathbf{m}}_i) \le \frac{1}{n}\cdot 2^{O(\ell\log\ell)}.$$

5.3. Full Algorithm and Main Theorem

In this subsection, we will present the full algorithm and combine with the analysis in the previous subsections to prove our main theorem.

Proof of Theorem 2.1. Suppose we are given n iid samples $x_1, \dots, x_n$ from the distribution (1). We will show that the estimated weights $\widehat{w}_1, \widehat{w}_2, \dots, \widehat{w}_k$ and the estimated means $\widehat{\mu}_1, \widehat{\mu}_2, \cdots, \widehat{\mu}_k$ outputted by Algorithm 1 taking $x_1, \dots, x_n$ as the input satisfy the desired guarantees.

By Lemma 5.2, when $\ell = \Omega(\frac{\log \frac{1}{\delta}}{\log\log \frac{1}{\delta}})$, we have

$$|\widehat{\mathbf{m}}_i - \mathbf{m}_i| < 2\delta \qquad \text{with probability } 1 - \frac{\mathsf{Var}(\widehat{\mathbf{m}}_i)}{\delta^2}$$

where $\widehat{\mathbf{m}}_i$ are computed in Algorithm 1. Moreover, by Lemma 5.3, we show that

$$\mathsf{Var}(\widehat{\mathbf{m}}_i) \le \frac{1}{n} \cdot 2^{O(\ell \log \ell)}$$

which implies when $\ell = \Omega(\frac{\log \frac{1}{\delta}}{\log\log \frac{1}{\delta}})$ the failure probability is less than

$$\frac{\mathsf{Var}(\widehat{\mathbf{m}}_i)}{\delta^2} \le \frac{1}{n} \cdot \mathsf{poly}(\frac{1}{\delta}).$$

By applying the union bound over all $i = 1, 2, \dots, 2k-1$, when $n = \Omega(\mathsf{poly}(\frac{1}{\delta}))$, we have

$$|\widehat{\mathbf{m}}_i - \mathbf{m}_i| < 2\delta \qquad \text{with probability } \frac{99}{100}.$$

In (Wu & Yang, 2018), they showed that Algorithm 2 is the algorithm that makes the guarantees hold in Theorem 5.1. Therefore, if we pick $\delta = \varepsilon^{\Omega(k)}$ along with the assumption $w_{\min}\Delta_{\min} = \Omega(\varepsilon)$, we have, up to an index permutation Π,

$$|\widehat{w}_{\Pi(i)} - w_i| < \varepsilon, \qquad |\widehat{\mu}_{\Pi(i)} - \mu_i| < \varepsilon \qquad \text{for } i = 1, \dots, k.$$

We now examine the running time of Algorithm 1. It first takes $k \cdot \mathsf{poly}(\ell)$ time[2] to obtain β_i. Then, it takes $n \cdot k \cdot \mathsf{poly}(\ell)$ to compute $\widehat{\mathbf{m}}_i$. Finally, the running time for Algorithm 2 is $\mathsf{poly}(k)$. Hence, by plugging $\ell = O(k \log \frac{1}{\varepsilon})$, the running time of Algorithm 1 is $n \cdot \mathsf{poly}(k, \frac{1}{\varepsilon})$.

□

Algorithm 2 Denoised method of moments (Wu & Yang, 2018)

```
Input: estimated moments $\widehat{\mathbf{m}} = (\widehat{\mathbf{m}}_1, \dots, \widehat{\mathbf{m}}_{2k-1})$, mean boundary parameter $M$
1: let $\mathbf{m}^* = (\mathbf{m}^*_1, \dots, \mathbf{m}^*_{2k-1})$ be the optimal solution of the following convex optimization problem
       $\arg\max_{\mathbf{m}} \|\widehat{\mathbf{m}} - \mathbf{m}\|$
       s.t. $M \cdot \mathbf{M}_{0,2k-2} \succcurlyeq \mathbf{M}_{1,2k-1} \succcurlyeq -M \cdot \mathbf{M}_{0,2k-2}$
   where $\mathbf{M}_{i,j}$ is the Hankel matrix whose entries are $\mathbf{m}_i, \dots, \mathbf{m}_j$, i.e.
       $\mathbf{M}_{i,j} = \begin{bmatrix} \mathbf{m}_i & \mathbf{m}_{i+1} & \cdots & \mathbf{m}_{\frac{i+j}{2}} \\ \mathbf{m}_{i+1} & \mathbf{m}_{i+2} & \cdots & \mathbf{m}_i \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m}_{\frac{i+j}{2}} & \mathbf{m}_{\frac{i+j}{2}+1} & \cdots & \mathbf{m}_j \end{bmatrix}$
2: let $\widehat{\mu}_1, \widehat{\mu}_2, \cdots, \widehat{\mu}_k$ be the roots of the polynomial $P$ where
       $P(x) = \det \begin{bmatrix} 1 & \mathbf{m}^*_1 & \cdots & \mathbf{m}^*_k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m}^*_{k-1} & \mathbf{m}^*_k & \cdots & \mathbf{m}^*_{2k-1} \\ 1 & x & \cdots & x^k \end{bmatrix}$
3: let $(\widehat{w}_1, \widehat{w}_2, \dots, \widehat{w}_k)^\top$ be the solution of the following system of linear equations
       $\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \widehat{\mu}_1 & \widehat{\mu}_2 & \cdots & \widehat{\mu}_k \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\mu}_1^{k-1} & \widehat{\mu}_2^{k-1} & \cdots & \widehat{\mu}_k^{k-1} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{m}^*_1 \\ \vdots \\ \mathbf{m}^*_{k-1} \end{bmatrix}$
Output: estimated weights $\widehat{w}_1, \widehat{w}_2, \dots, \widehat{w}_k$ and estimated means $\widehat{\mu}_1, \widehat{\mu}_2, \cdots, \widehat{\mu}_k$
```

6. Conclusion and Discussion

In this paper, we study the classical problem of learning mixtures of Gaussians with censored data. The problem becomes more challenging compared to the problem of learning with uncensored data because the data are partially observed. Our result shows that there is an efficient algorithm to estimate the weights and the means of the Gaussians. Specifically, we show that one only needs $\frac{1}{\varepsilon^{O(k)}}$ censored samples to estimate the weights and the means within ε error. To the best of our knowledge, this is the first finite sample bound for the problem of learning mixtures of Gaussians with censored data even in the simple setting that the Gaussians are univariate and homogeneous.

There are multiple natural extensions to this setting. For example, a natural extension is to consider mixtures of multivariate Gaussians. Without truncation or censoring, one popular approach to learn mixtures of multivariate Gaus-

[2] Computing the integral in (15) can be reduced to computing the integral $\int_0^z e^{-\frac{1}{2}t^2} \mathrm{d}t$ by observing h_c and h_r are polynomials and using integration by parts. If we remove the assumption that the exact computation can be done, we will need to approximate the integral up to an additive error of $1/2^{\mathsf{poly}(k,\ell)}$. One can approximate the integral in an exponential convergence rate by Taylor expansion and hence the running time is still $k \cdot \mathsf{poly}(\ell)$ for this step.

sians is to apply random projections and reduce the problem to univariate Gaussians. This approach relies on the fact that the projection of a mixture of Gaussians is also a mixture of Gaussians. Unfortunately, this fact is no longer true when the data are truncated or censored.

Another interesting direction is to relax the assumption of known and homogeneous variances to unknown and/or non-homogeneous variances. When the Gaussians are homogeneous, one can estimate the variance by computing the pairwise distances between $k + 1$ samples and find the minimum of them if the samples are not truncated or censored. It holds from the fact that two samples are from the same Gaussian and hence the expected value of their squared distance is the variance. It becomes more challenging when the samples are truncated or censored because the expected value of the squared distance may not be the variance.

Furthermore, previous results indicate that, in the uncensored setting, sample bounds can be improved when the centers of Gaussians in the mixture are well-separated (Moitra, 2015; Regev & Vijayaraghavan, 2017; Qiao et al., 2022). An interesting direction for future research would be to improve our results under stronger separation assumptions on the components. For example, one strategy to exploit separation is to apply the Fourier Transform to the pdf of the mixture. With uncensored samples, it is straightforward to estimate the Fourier Transform, however, when the pdf is truncated, a challenge arises as the Fourier Transform may not yield a convenient form, as required by these analyses. We anticipate that delicate modifications may still be needed, and leave this open to future work.

References

Amemiya, T. Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pp. 997–1016, 1973.

Balakrishnan, N. and Cramer, E. The art of progressive censoring. *Statistics for industry and technology*, 2014.

Bernoulli, D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, pp. 1–45, 1760.

Cohen, A. C. *Truncated and Censored Samples: Theory and Applications*. CRC Press, 2016.

Dasgupta, S. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pp. 634–644. IEEE, 1999.

Daskalakis, C., Tzamos, C., and Zampetakis, M. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pp. 704–710. PMLR, 2017.

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 639–649. IEEE, 2018.

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory*, pp. 955–960. PMLR, 2019.

Diakonikolas, I. and Kane, D. M. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pp. 999–1008. PMLR, 2017.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2683–2702. SIAM, 2018.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Doss, N., Wu, Y., Yang, P., and Zhou, H. H. Optimal estimation of high-dimensional gaussian mixtures. *arXiv preprint arXiv:2002.05818*, 2020.

Fisher, R. Properties and applications of hh functions. *Mathematical tables*, 1:815–852, 1931.

Galton, F. An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315, 1898.

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

Hausman, J. A. and Wise, D. A. Social experimentation, truncated distributions, and efficient estimation. *Econometrica: Journal of the Econometric Society*, pp. 919–938, 1977.

Hopkins, S. B., Kamath, G., and Majid, M. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1406–1417, 2022.

Kalai, A. T., Moitra, A., and Valiant, G. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 553–562, 2010.

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

Lee, A. Table of the gaussian" tail" functions; when the" tail" is larger than the body. *Biometrika*, 10(2/3):208–214, 1914.

Lee, G. and Scott, C. Em algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816–2829, 2012.

Lindsay, B. G. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–163. JSTOR, 1995.

Liu, X., Kong, W., Kakade, S., and Oh, S. Robust and differentially private mean estimation. *Advances in neural information processing systems*, 34:3887–3901, 2021.

Maddala, G. S. *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press, 1986.

McLachlan, G. and Jones, P. Fitting mixture models to grouped and truncated data via the em algorithm. *Biometrics*, pp. 571–578, 1988.

Moitra, A. Super-resolution, extremal functions and the condition number of vandermonde matrices. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 821–830, 2015.

Moitra, A. and Valiant, G. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 93–102. IEEE, 2010.

Nagarajan, S. G. and Panageas, I. On the analysis of em for truncated mixtures of two gaussians. In *Algorithmic Learning Theory*, pp. 634–659. PMLR, 2020.

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

Pearson, K. On the systematic fitting of curves to observations and measurements. *Biometrika*, 1(3):265–303, 1902.

Pearson, K. and Lee, A. On the generalised probable error in multiple normal correlation. *Biometrika*, 6(1):59–68, 1908.

Qiao, M., Guruganesh, G., Rawat, A., Dubey, K. A., and Zaheer, M. A fourier approach to mixture learning. *Advances in Neural Information Processing Systems*, 35: 20850–20861, 2022.

Regev, O. and Vijayaraghavan, A. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 85–96. IEEE, 2017.

Schneider, H. *Truncated and censored samples from normal populations*. Marcel Dekker, Inc., 1986.

Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pp. 24–36, 1958.

Vempala, S. and Wang, G. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

Wu, Y. and Yang, P. Optimal estimation of gaussian mixtures via denoised method of moments. *arXiv preprint arXiv:1807.07237*, 2018.

Xu, J., Hsu, D. J., and Maleki, A. Global analysis of expectation maximization for mixtures of two gaussians. *Advances in Neural Information Processing Systems*, 29, 2016.

A. Proof

In this section, we will present the proofs of the lemmas.

Lemma A.1. *Let V be the matrix defined in* (8)*, i.e. V is the ℓ-by-ℓ matrix whose (r,c)-entry is $J_{h_c,r}$ for $r,c = 0,1,\dots,\ell-1$. Recall that, from* (5)*, $J_{h_c,r}$ is defined as*

$$J_{h_c,r} = \int_S \frac{1}{\sqrt{2\pi} r!} e^{-\frac{1}{2}x^2} h_c(x) h_r(x) \mathsf{d}x.$$

Then, the determinant of V is

$$\det(V) = \left(\frac{1}{\sqrt{2\pi}}\right)^{\ell} \cdot \prod_{r=0}^{\ell-1} \frac{1}{r!} \cdot \int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x} \in S^\ell} e^{-\frac{1}{2}\sum_{c=0}^{\ell-1} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-1} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x}.$$

Proof. Since the (r,c)-entry of V is

$$J_{h_c,r} = \int_S \frac{1}{\sqrt{2\pi} r!} e^{-\frac{1}{2}x^2} h_c(x) h_r(x) \mathsf{d}x,$$

by factoring out the term $\frac{1}{\sqrt{2\pi} r!}$ for each row, we have

$$\det(V) = \left(\frac{1}{\sqrt{2\pi}}\right)^{\ell} \cdot \prod_{r=0}^{\ell-1} \frac{1}{r!} \cdot \det(W) \tag{16}$$

where W is the ℓ-by-ℓ matrix whose (r,c)-entry is

$$W_{r,c} = \int_S e^{-\frac{1}{2}x^2} h_c(x) h_r(x) \mathsf{d}x. \tag{17}$$

By Cauchy-Binet formula, we can further express $\det(W)$ as

$$\det(W) = \int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x} \in S^\ell} (\det(U(\mathbf{x})))^2 \mathsf{d}\mathbf{x} \tag{18}$$

where $U(\mathbf{x})$ is the ℓ-by-ℓ matrix whose (r,c)-entry is

$$U(\mathbf{x})_{r,c} = e^{-\frac{1}{4}x_c^2} h_r(x_c) \tag{19}$$

for any $\mathbf{x} = (x_0, \dots, x_{\ell-1}) \in S^\ell$. By factoring out the term $e^{-\frac{1}{4}x_c^2}$ for each column, we have

$$\det(U(\mathbf{x})) = e^{-\frac{1}{4}\sum_{c=0}^{\ell-1} x_c^2} \det(P(\mathbf{x})) \tag{20}$$

where $P(\mathbf{x})$ is the ℓ-by-ℓ matrix whose (r,c)-entry is

$$P(\mathbf{x})_{r,c} = h_r(x_c) \tag{21}$$

for any $\mathbf{x} = (x_0, \dots, x_{\ell-1}) \in S^\ell$. Since h_r is a polynomial of degree r with the leading coefficient 1, by applying row and column operations, the determinant $\det(P(\mathbf{x}))$ is same as the determinant of the Vandermonde matrix, i.e.

$$\det(P(\mathbf{x})) = \prod_{0 \le c_1 < c_2 \le \ell-1} (x_{c_1} - x_{c_2}). \tag{22}$$

In other words, the determinant $\det(V)$ is

$$\det(V) = \left(\frac{1}{\sqrt{2\pi}}\right)^{\ell} \cdot \prod_{r=0}^{\ell-1} \frac{1}{r!} \cdot \int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x} \in S^\ell} e^{-\frac{1}{2}\sum_{c=0}^{\ell-1} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-1} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x}.$$

□

Lemma A.2. *Let $V^{(i\to j)}$ be the matrix defined in (10) for $i \le 2k-1$ and $j \ge \ell \ge 2(2k-1) \ge 2i$. Then the absolute value of the determinant of $V^{(i\to j)}$ is*

$$|\det(V^{(i\to j)})| \le \frac{1}{2^{\Omega(j\log j)}} \cdot |\det(V)|.$$

Proof. We can perform a similar computation as in the computation of $\det(V)$. Namely, we factor out the term $\frac{1}{\sqrt{2\pi r!}}$ for each row, we have

$$|\det(V^{(i\to j)})| = \left(\frac{1}{\sqrt{2\pi}}\right)^{\ell} \cdot \prod_{r=0, r\neq i}^{\ell-1} \frac{1}{r!} \cdot \frac{1}{j!} \cdot |\det(W^{(i\to j)})|$$

where $W^{(i\to j)}$ is the same matrix as W from (17) except that the i-th row is replaced by the row $\sqrt{2\pi} j! v^{(j)}$. By comparing to (16), we simplify $|\det(V^{(i\to j)})|$ to be

$$|\det(V^{(i\to j)})| = \frac{i!}{j!} \cdot \frac{|\det(W^{(i\to j)})|}{|\det(W)|} \cdot |\det(V)| \tag{23}$$

By Cauchy-Binet formula, we can further express $\det(W^{(i\to j)})$ as

$$\det(W^{(i\to j)}) = \int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x}\in S^\ell} \det(U(\mathbf{x})) \det(U^{(i\to j)}(\mathbf{x})) \mathsf{d}\mathbf{x}$$

where $U^{(i\to j)}(\mathbf{x})$ is the same matrix as $U(\mathbf{x})$ from (19) except that the i-th row is replaced with the column whose c-th entry is $e^{-\frac{1}{4}x_c^2} h_j(x_c)$ for any $\mathbf{x} = (x_0, \dots, x_{\ell-1}) \in \mathbb{R}^\ell$. Furthermore, by Cauchy–Schwarz inequality and comparing to (18),

$$\begin{aligned} |\det(W^{(i\to j)})| &\le \left(\int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x}\in S^\ell} (\det(U(\mathbf{x})))^2 \mathsf{d}\mathbf{x}\right)^{1/2} \left(\int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x}\in S^\ell} (\det(U^{(i\to j)}(\mathbf{x})))^2 \mathsf{d}\mathbf{x}\right)^{1/2} \\ &= \left(\frac{\int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x}\in S^\ell} (\det(U^{(i\to j)}(\mathbf{x})))^2 \mathsf{d}\mathbf{x}}{\int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x}\in S^\ell} (\det(U(\mathbf{x})))^2 \mathsf{d}\mathbf{x}}\right)^{1/2} |\det(W)|. \end{aligned} \tag{24}$$

By factoring out the term $e^{-\frac{1}{4}x_c^2}$ for each column, we have

$$\det(U^{(i\to j)}(\mathbf{x})) = e^{-\frac{1}{4}\sum_{c=0}^{\ell-1} x_c^2} \det(P^{(i\to j)}(\mathbf{x})) \tag{25}$$

where $P^{(i\to j)}(\mathbf{x})$ is the same matrix as $P(\mathbf{x})$ from (21) except that the i-th row is replaced with the row whose c-th entry is $h_j(x_c)$ for any $\mathbf{x} = (x_0, \dots, x_{\ell-1}) \in \mathbb{R}^\ell$.

This time, the computation of $\det(P^{(i\to j)}(\mathbf{x}))$ is not as easy as $\det(P(\mathbf{x}))$. In Lemma A.4 below, we will show that

$$|\det(P^{(i\to j)}(\mathbf{x}))| \le \frac{j!}{i!(\frac{j-i}{2})!} \cdot 2^{O(j)} \cdot |\det(P(\mathbf{x}))|.$$

Plugging it into (25) and comparing (25) to (20), we have

$$|\det(U^{(i\to j)}(\mathbf{x}))| \le \frac{j!}{i!(\frac{j-i}{2})!} \cdot 2^{O(j)} \cdot |\det(U(\mathbf{x}))|.$$

Furthermore, by plugging it into (24),

$$|\det(W^{(i\to j)})| \le \left(\frac{\int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x}\in S^\ell} (\det(U^{(i\to j)}(\mathbf{x})))^2 \mathsf{d}\mathbf{x}}{\int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x}\in S^\ell} (\det(U(\mathbf{x})))^2 \mathsf{d}\mathbf{x}}\right)^{1/2} |\det(W)| \le \frac{j!}{i!(\frac{j-i}{2})!} \cdot 2^{O(j)} \cdot |\det(W)|$$

Finally, when we plug it into (23), we prove that

$$|\det(V^{(i\to j)})| = \frac{i!}{j!}\cdot\frac{|\det(W^{(i\to j)})|}{|\det(W)|}\cdot|\det(V)| \le \frac{i!}{j!}\cdot\frac{j!2^{O(j)}}{i!(\frac{j-i}{2})!}\cdot|\det(V)| = \frac{2^{O(j)}}{(\frac{j-i}{2})!}\cdot|\det(V)|$$

Recall that $i \le 2k-1$ and the assumption of $j \ge \ell > 2(2k-1) \ge 2i$. We have

$$|\det(V^{(i\to j)})| \le \frac{1}{2^{\Omega(j\log j)}}\cdot|\det(V)|.$$

□

Lemma A.3. *Let $V^{(\mathbf{e}_i\leftarrow a)}$ be the matrix defined in (13) for $i \le 2k-1$ and $a \le \ell$. Then the absolute value of the determinant of $V^{(\mathbf{e}_i\leftarrow a)}$ is*

$$|\det(V^{(\mathbf{e}_i\leftarrow a)})| \le 2^{O(\ell\log\ell)}\cdot|\det(V)|.$$

Proof. Recall that $V^{(\mathbf{e}_i\leftarrow a)}$ is the same matrix as V except that the a-th column is replaced with $\mathbf{e}_i$. Hence, we first expand the determinant along that column and factor out the term $\frac{1}{\sqrt{2\pi}r!}$ for each row.

$$|\det(V^{(\mathbf{e}_i\leftarrow a)})| = \left(\frac{1}{\sqrt{2\pi}}\right)^{\ell-1}\cdot\prod_{r=0,r\ne i}^{\ell-1}\frac{1}{r!}\cdot|\det(W^{(-i,-a)})|$$

where $W^{(-i,-a)}$ is the same matrix as W from (17) except that the i-th row and the a-th column are omitted. By comparing to (16), we first simplify $|\det(V^{(\mathbf{e}_i\leftarrow a)})|$ to be

$$|\det(V^{(\mathbf{e}_i\leftarrow a)})| = \sqrt{2\pi}i!\cdot\frac{|\det(W^{(-i,-a)})|}{|\det(W)|}\cdot|\det(V)|$$

It means we need to bound the term $\frac{|\det(W^{(-i,-a)})|}{|\det(W)|}$ from above. To achieve it, we will bound $|\det(W^{(-i,-a)})|$ from above and $|\det(W)|$ from below.

By Cauchy-Binet formula, we further express $\det(W^{(-i,-a)})$ as

$$\det(W^{(-i,-a)}) = \int_{x_0>\cdots>x_{\ell-2},\mathbf{x}\in S^{\ell-1}}\det(U^{(-i)}(\mathbf{x}))\det(U^{(-a)}(\mathbf{x}))\mathsf{d}\mathbf{x} \tag{26}$$

where $U^{(-i)}(\mathbf{x})$ (resp. $U^{(-a)}$) is the $(\ell-1)$-by-$(\ell-1)$ matrix whose (r,c)-entry is $e^{-\frac{1}{4}x_c^2}h_r(x_c)$ for $r \in [\ell]\backslash\{i\}$ (resp. $r \in [\ell]\backslash\{a\}$), $c \in [\ell-1]$ and any $\mathbf{x} = (x_0,\dots,x_{\ell-2}) \in \mathbb{R}^{\ell-1}$. By factoring out the term $e^{-\frac{1}{4}x_c^2}$ fro each column,

$$\det(U^{(-i)}(\mathbf{x})) = e^{-\frac{1}{4}\sum_{c=0}^{\ell-2}x_c^2}\det(P^{(-i)}(\mathbf{x})) \tag{27}$$

where $P^{(-i)}(\mathbf{x})$ is the $(\ell-1)$-by-$(\ell-1)$ matrix whose (r,c)-entry is $h_r(x_c)$ for $r \in [\ell]\backslash\{i\}$, $c \in [\ell-1]$ and any $\mathbf{x} = (x_0,\dots,x_{\ell-2}) \in \mathbb{R}^{\ell-1}$.

Again, the computation of $\det(P^{(-i)}(\mathbf{x}))$ is not as easy as $\det(P(\mathbf{x}))$. In Lemma A.5, we show that

$$|\det(P^{(-i)}(\mathbf{x}))| \le 2^{O(\ell\log\ell)}\cdot\prod_{1\le c_1<c_2\le\ell-2}|x_{c_1}-x_{c_2}|.$$

Note that the bound is independent to i and hence we have the same bound for $|P^{(-a)}(\mathbf{x})|$. By plugging it into (27) and further into (26), we have

$$|\det(W^{(-i,-a)})| \le 2^{O(\ell\log\ell)}\cdot\int_{x_0>\cdots>x_{\ell-2},\mathbf{x}\in S^{\ell-1}}e^{-\frac{1}{2}\sum_{c=0}^{\ell-2}x_c^2}\cdot\prod_{1\le c_1<c_2\le\ell-2}(x_{c_1}-x_{c_2})^2\mathsf{d}\mathbf{x}. \tag{28}$$

Recall that, in Lemma A.1 and (16),

$$\det(W) = \int_{x_0 > \cdots > x_{\ell-1}, \mathbf{x} \in S^\ell} e^{-\frac{1}{2}\sum_{c=0}^{\ell-1} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-1} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x}.$$

Since the term $e^{-\frac{1}{2}\sum_{c=0}^{\ell-1} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-1} (x_{c_1} - x_{c_2})^2$ in the integral is symmetric with respect to $x_0, \dots, x_{\ell-1}$, we have

$$\det(W) = \ell! \cdot \int_{\mathbf{x} \in S^\ell} e^{-\frac{1}{2}\sum_{c=0}^{\ell-1} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-1} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x}.$$

To bound $\det(W)$ from below, we consider integrating over the sub-region $\left\{\mathbf{x} \in S^\ell \mid |x_{\ell-1} - x_c| > \frac{R}{\ell}\right\}$ of S^ℓ.

$$\begin{aligned}
\det(W) &\ge \ell! \cdot \int_{|x_{\ell-1} - x_c| > \frac{R}{\ell}, \mathbf{x} \in S^\ell} e^{-\frac{1}{2}\sum_{c=0}^{\ell-1} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-1} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x} \\
&\ge \ell! \cdot \left(\frac{R}{\ell}\right)^{2(\ell-1)} e^{-\frac{1}{2}R^2} \cdot \int_{|x_{\ell-1} - x_c| > \frac{R}{\ell}, \mathbf{x} \in S^\ell} e^{-\frac{1}{2}\sum_{c=0}^{\ell-2} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-2} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x} \\
&\ge \ell! \cdot R \left(\frac{R}{\ell}\right)^{2(\ell-1)} e^{-\frac{1}{2}R^2} \cdot \int_{\mathbf{x} \in S^{\ell-1}} e^{-\frac{1}{2}\sum_{c=0}^{\ell-2} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-2} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x} \\
&= \ell \cdot R \left(\frac{R}{\ell}\right)^{2(\ell-1)} e^{-\frac{1}{2}R^2} \cdot \int_{x_0 > \cdots > x_{\ell-2}, \mathbf{x} \in S^{\ell-1}} e^{-\frac{1}{2}\sum_{c=0}^{\ell-2} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-2} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x} \\
&= \frac{1}{2^{O(\ell \log \ell)}} \cdot \int_{x_0 > \cdots > x_{\ell-2}, \mathbf{x} \in S^{\ell-1}} e^{-\frac{1}{2}\sum_{c=0}^{\ell-2} x_c^2} \cdot \prod_{0 \le c_1 < c_2 \le \ell-2} (x_{c_1} - x_{c_2})^2 \mathsf{d}\mathbf{x} \qquad (29)
\end{aligned}$$

In other words, by comparing $|\det(W)|$ in (29) to $|\det(W^{(-i,-a)})|$ in (28), we have

$$\frac{|\det(W^{(-i,-a)})|}{|\det(W)|} \le 2^{O(\ell \log \ell)}$$

and hence

$$\frac{|\det(V^{(\mathbf{e}_i \leftarrow a)})|}{|\det(V)|} = \sqrt{2\pi} i! \cdot \frac{|\det(W^{(-i,-a)})|}{|\det(W)|} \le 2^{O(\ell \log \ell)}.$$

□

Lemma A.4. *Let $P^{(i \to j)}(\mathbf{x})$ be the matrix defined in the proof of Lemma A.2. Then the absolute value of the determinant of $P^{(i \to j)}(\mathbf{x})$ is*

$$|\det(P^{(i \to j)}(\mathbf{x}))| \le \frac{j!}{i!(\frac{j-i}{2})!} \cdot 2^{O(j)} \cdot |\det(P(\mathbf{x}))|.$$

Recall that $P(\mathbf{x})$ is the matrix defined in (21).

Proof. Since the entries of $P^{(i \to j)}(\mathbf{x})$ are Hermite polynomials, we can decompose it into

$$P^{(i \to j)}(\mathbf{x}) = C^{(i \to j)} \cdot X^{[j+1]}$$

where $C^{(i \to j)}$ is the ℓ-by-$(j+1)$ matrix whose (r, c)-entry is the coefficient of x^c in the r-th Hermite polynomial and $X^{[j+1]}$ is the $(j+1)$-by-ℓ matrix whose (r, c)-entry is x_c^r. For example, take $\ell = 4, i = 2, j = 6$,

$$\begin{aligned}
h_0(x) &= 1 \\
h_1(x) &= x \\
h_3(x) &= -3x + x^3 \\
h_6(x) &= -15 + 45x^2 - 15x^4 + x^6
\end{aligned}$$

and hence

$$C^{(i\to j)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -15 & 0 & 45 & 0 & -15 & 0 & 1 \\ 0 & -3 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

To compute $\det(P^{(i\to j)}(\mathbf{x}))$, we use Cauchy-Binet formula and we have

$$\det(P^{(i\to j)}(\mathbf{x})) = \sum_T \det(C^{(i\to j)}_{:,T}) \cdot \det(X^{[j+1]}_{T,:})$$

where the summation is over all subset T of size ℓ of $[j+1]$, $C^{(i\to j)}_{:,T}$ is the ℓ-by-ℓ matrix whose columns are the columns of $C^{(i\to j)}$ at indices from T and $X^{[j+1]}_{T,:}$ is the ℓ-by-ℓ matrix whose rows are the rows of $X^{[j+1]}$ at indices from T. Here, for any positive integer n, we denote $[n]$ to be the set $\{0, 1, \ldots, n-1\}$. Furthermore, by triangle inequality,

$$|\det(P^{(i\to j)}(\mathbf{x}))| \le \sum_T |\det(C^{(i\to j)}_{:,T})| \cdot |\det(X^{[j+1]}_{T,:})| \tag{30}$$

We first make some simplifications to see what T makes the determinants nonzero. For example, take $\ell = 8, i = 2, j = 10$, we have

$$\begin{aligned} h_0(x) &= 1 \\ h_1(x) &= x \\ h_3(x) &= -3x + x^3 \\ h_4(x) &= 3 - 6x^2 + x^4 \\ h_5(x) &= 15x - 10x^3 + x^5 \\ h_6(x) &= -15 + 45x^2 - 15x^4 + x^6 \\ h_7(x) &= -105x + 105x^3 - 21x^5 + x^7 \\ h_{10}(x) &= -945 + 4725x^2 - 3150x^4 + 630x^6 - 45x^8 + x^{10} \end{aligned}$$

and

$$C^{(i,j)} =_{\text{up to row and column swaps}} \begin{bmatrix} 1 & & & & & & & & & \\ 3 & -6 & 1 & & & & & & & \\ -15 & 45 & -15 & 1 & & & & & & \\ -945 & 4725 & -3150 & 630 & -45 & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & -3 & 1 & & \\ & & & & & & 15 & -10 & 1 & \\ & & & & & & -105 & 105 & -21 & 1 \end{bmatrix}$$

For simplicity, we assume that i, j, ℓ are even numbers and it is easy to prove the other cases by symmetry. If T satisfies one of the following conditions:

- does not contain all odd numbers less than ℓ, i.e. $1, 3, \ldots, \ell - 1$
- does not contain all even numbers less than i, i.e. $0, 2, \ldots, i - 2$
- contains more than one even number larger than or equal to ℓ, i.e. $\ell, \ell + 2, \ldots, j$

then $\det(C^{(i\to j)}_{:,T}) = 0$. In other words, the choices are

- $T = [\ell]$ or

- $T = [\ell] \backslash \{a\} \cup \{b\}$ for $a = i, i+2, \ldots, \ell-2$ and $b = \ell, \ell+2, \ldots, j$.

Therefore, there are only $\frac{\ell-i}{2} \cdot \frac{j-\ell+2}{2} + 1 = O(j^2)$ choices for T such that $\det(C^{(i\to j)}_{:,T})$ may not be 0.

If $T = [\ell]$, by expanding the determinant $\det(C^{(i\to j)}_{:,T})$ along the rows whose diagonal entry is 1, what we have left is the determinant of a matrix A where A is the $(\frac{\ell-i}{2})$-by-$(\frac{\ell-i}{2})$ matrix whose (r,c)-entry is $(-1)^{\frac{r-c}{2}} \frac{r!}{(\frac{r-c}{2})!c!2^{\frac{r-c}{2}}}$ for $r = i+2, \ldots, \ell-2, j$ and $c = i, i+2, \ldots, \ell-2$. In the example, the matrix A is $\begin{bmatrix} -6 & 1 & \\ 45 & -15 & 1 \\ 4725 & -3150 & 630 \end{bmatrix}$. By applying row and column operations, we can compute the exact expression for $\det(A)$

$$\det(A) = (-1)^{\frac{j-i}{2}} \frac{j!}{i! 2^{\frac{j-i}{2}}} \left(\sum_{m=0}^{\frac{\ell-i-2}{2}} (-1)^m \frac{1}{m!(\frac{j-i}{2}-m)!} \right).$$

In the example, we have

$$\det(\begin{bmatrix} -6 & 1 & \\ 45 & -15 & 1 \\ 4725 & -3150 & 630 \end{bmatrix}) = 14175$$

Note that the expression $\sum_{m=0}^{\frac{\ell-i-2}{2}} (-1)^m \frac{1}{m!(\frac{j-i}{2}-m)!}$ in the equation for $\det(A)$ can be easily bounded by

$$|\sum_{m=0}^{\frac{\ell-i-2}{2}} (-1)^m \frac{1}{m!(\frac{j-i}{2}-m)!}| \le \sum_{m=0}^{\frac{\ell-i-2}{2}} \frac{1}{m!(\frac{j-i}{2}-m)!} \le \sum_{m=0}^{\frac{j-i}{2}} \frac{1}{m!(\frac{j-i}{2}-m)!} = \frac{2^{\frac{j-i}{2}}}{(\frac{j-i}{2})!}$$

Hence, we have

$$|\det(C^{(i\to j)}_{:,T})| = |\det(A)| \le \frac{j!}{i!(\frac{j-i}{2})!}$$

Also, since $T = [\ell]$, therefore $|\det(X^{[j+1]}_{T,:})| = \prod_{0 \le c_1 < c_2 \le \ell-1} |x_{c_1} - x_{c_2}|$. When $T = [\ell]$, we have

$$|\det(C^{(i\to j)}_{:,T})| \cdot |\det(X^{[j+1]}_{T,:})| \le \frac{j!}{i!(\frac{j-i}{2})!} \cdot \prod_{0 \le c_1 < c_2 \le \ell-1} |x_{c_1} - x_{c_2}|$$

Now, consider the case that $T = [\ell] \backslash \{a\} \cup \{b\}$ for $a = i, i+2, \ldots, \ell-2$ and $b = \ell, \ell+2, \ldots, j$. Similar to the previous calculation, by expanding the determinant $\det(C^{(i\to j)}_{:,T})$ along the rows whose diagonal entry is 1, what we have left is the determinant of a matrix A where A is the $(\frac{\ell-i}{2})$-by-$(\frac{\ell-i}{2})$ matrix whose (r,c)-entry is $(-1)^{\frac{r-c}{2}} \frac{r!}{(\frac{r-c}{2})!c!2^{\frac{r-c}{2}}}$ for $r = i+2, \ldots, a, j$ and $c = i, i+2, \ldots, a-2, b$. For example, take $a = 6$ and $b = 8$, the matrix A is the example is $\begin{bmatrix} -6 & 1 & \\ 45 & -15 & \\ 4725 & -3150 & -45 \end{bmatrix}$. By applying row and column operations, we can compute the exact expression for $\det(A)$

$$\det(A) = (-1)^{\frac{j-b}{2}} \frac{j!}{(\frac{j-b}{2})!b!2^{\frac{j-b}{2}}} \cdot (-1)^{\frac{a-i}{2}} \frac{a!}{(\frac{a-i}{2})!i!2^{\frac{a-i}{2}}}$$

In the example, we have

$$\det(\begin{bmatrix} -6 & 1 & \\ 45 & -15 & \\ 4725 & -3150 & -45 \end{bmatrix}) = -2025$$

To bound $|\det(A)|$,

$$|\det(A)| = \frac{j!}{(\frac{j-b}{2})!b!2^{\frac{j-b}{2}}} \cdot \frac{a!}{(\frac{a-i}{2})!i!2^{\frac{a-i}{2}}} = \frac{j!}{i!} \cdot \frac{a!}{(\frac{a-i}{2})!b!(\frac{j-b}{2})!} \cdot \frac{1}{2^{\frac{j-b+a-i}{2}}}$$

Note that $\frac{1}{2^{\frac{j-b+a-i}{2}}} \leq 1$. Recall that $i \leq a \leq \ell - 2$ and $\ell \leq b \leq j$. We also have

$$\frac{a!}{(\frac{a-i}{2})!} \leq 2^a \cdot (\frac{a+i}{2})! \leq 2^j \cdot (\frac{b+i}{2})!.$$

Hence,

$$|\det(A)| \leq \frac{j!}{i!} \cdot \frac{2^j(\frac{b+i}{2})!}{b!(\frac{j-b}{2})!} = \frac{j!}{i!} \cdot 2^j \cdot \frac{(\frac{b+i}{2})!(\frac{b-i}{2})!}{b!} \cdot \frac{(\frac{j-i}{2})!}{(\frac{b-i}{2})!(\frac{j-b}{2})!} \cdot \frac{1}{(\frac{j-i}{2})!}$$

Observe that

$$\frac{(\frac{b+i}{2})!(\frac{b-i}{2})!}{b!} \leq 1 \qquad \text{and} \qquad \frac{(\frac{j-i}{2})!}{(\frac{b-i}{2})!(\frac{j-b}{2})!} \leq 2^{\frac{j-i}{2}} \leq 2^{\frac{j}{2}}.$$

By plugging them into the above inequality,

$$|\det(C^{(i\to j)}_{:,T})| = |\det(A)| \leq \frac{j!}{i!} \cdot \frac{2^{\frac{3j}{2}}}{(\frac{j-i}{2})!}$$

Since a is omitted from $\{i, i+2, \ldots, \ell-2\}$ and b is selected from $\{\ell, \ell+2, \ldots, j\}$, it means that $T = [\ell]\backslash\{a\} \cup \{b\}$. By the properties of Schur polynomials,

$$\det(X^{[j+1]}_{T,:}) = \left(\sum_Y \mathbf{x}^Y\right) \cdot \prod_{1\leq c_1 < c_2 \leq \ell-1} (x_{c_1} - x_{c_2})$$

where the summation is over all semi-standard Young tableaux Y of shape $(b-\ell+1, \underbrace{1,\ldots,1}_{\ell-1-a \text{ 1's}}, \underbrace{0,\ldots,0}_{a \text{ 0's}})$. Here, the term $\mathbf{x}^Y$ means $x_0^{y_0} \cdots x_{\ell-1}^{y_{\ell-1}}$ where y_m is the number of occurrences of the number m in Y and note that $\sum_{m=0}^{\ell-1} y_m = b - a$. Based on the given shape, there is one row of size $b - \ell - 1$ and one column of size $\ell - a$ and they connect at the first element. For the row, the number of non-decreasing sequences of size $b - \ell - 1$ whose numbers are between 0 and $\ell - 1$ inclusive is $\binom{b}{\ell-1} \leq 2^j$. For the column, the number of increasing sequences of size $\ell - a$ whose numbers are between 0 and $\ell - 1$ inclusive is $\binom{\ell}{a} \leq 2^j$. Hence, the number of semi-standard Young tableaux of such shape is bounded by $\binom{b}{\ell-1} \cdot \binom{\ell}{a} \leq 2^{2j}$. By the assumption that $S = [-R, R]$, we can also bound the term $|\mathbf{x}^Y|$ to be

$$|\mathbf{x}^Y| \leq R^{b-a} \leq 2^{O(j)}.$$

We can now bound the determinant $|\det(X^{[j+1]}_{T,:})|$ by

$$|\det(X^{[j+1]}_{T,:})| \leq 2^{O(j)} \cdot \prod_{1\leq c_1 < c_2 \leq \ell-1} (x_{c_1} - x_{c_2}).$$

Namely, when $T = [\ell]\backslash\{a\} \cup \{b\}$ for $a = i, i+2, \ldots, \ell-2$ and $b = \ell, \ell+2, \ldots, j$,

$$\begin{aligned} |\det(C^{(i\to j)}_{:,T})| \cdot |\det(X^{[j+1]}_{T,:})| &\leq \frac{j!}{i!} \cdot \frac{2^{\frac{3j}{2}}}{(\frac{j-i}{2})!} \cdot 2^{O(j)} \cdot \prod_{1\leq c_1 < c_2 \leq \ell-1} (x_{c_1} - x_{c_2}) \\ &= \frac{j!}{i!(\frac{j-i}{2})!} \cdot 2^{O(j)} \cdot \prod_{0\leq c_1 < c_2 \leq \ell-1} |x_{c_1} - x_{c_2}| \end{aligned}$$

By considering all cases for T and plugging them into (30), we have

$$|\det(P^{(i\to j)}(\mathbf{x}))| \le \sum_T |\det(C^{(i\to j)}_{:,T})| \cdot |\det(X^{[j+1]}_{T,:})| \le \frac{j!}{i!(\frac{j-i}{2})!} \cdot 2^{O(j)} \cdot \prod_{0\le c_1<c_2\le \ell-1} |x_{c_1} - x_{c_2}|$$

and, by comparing to $\det(P(\mathbf{x}))$ in (22) which is $\prod_{0\le c_1<c_2\le\ell-1}|x_{c_1}-x_{c_2}|$,

$$|\det(P^{(i\to j)}(\mathbf{x}))| \le \frac{j!}{i!(\frac{j-i}{2})!} \cdot 2^{O(j)} \cdot |\det(P(\mathbf{x}))|.$$

□

Lemma A.5. *Let $P^{(-i)}(\mathbf{x})$ be the matrix defined in the proof of Lemma A.3. Then the absolute value of the determinant of $P^{(-i)}(\mathbf{x})$ is*

$$|\det(P^{(-i)}(\mathbf{x}))| \le 2^{O(\ell\log\ell)} \cdot \prod_{1\le c_1<c_2\le\ell-2} |x_{c_1} - x_{c_2}|.$$

Proof. Since the entries of $P^{(-i)}(\mathbf{x})$ are Hermite polynomials, we can decompose it into

$$P^{(-i)}(\mathbf{x}) = C^{(-i)} \cdot X^{[\ell]}$$

where $C^{(-i)}$ is the $(\ell-1)$-by-ℓ matrix whose (r,c)-entry is the coefficient of x^c in the r-th Hermite polynomial for $r\in[\ell]\backslash\{i\}$ and $X^{[\ell]}$ is the ℓ-by-$(\ell-1)$ matrix whose (r,c)-entry is x_c^r. For example, take $\ell=4, i=2$,

$$\begin{aligned} h_0(x) &= 1\\ h_1(x) &= x\\ h_3(x) &= -3x + x^3 \end{aligned}$$

and hence

$$C^{(-i)} = \begin{bmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & -3 & 0 & 1 \end{bmatrix} \qquad \text{and} \qquad X^{[\ell]} = \begin{bmatrix} 1 & 1 & 1\\ x_0 & x_1 & x_2\\ x_0^2 & x_1^2 & x_2^2\\ x_0^3 & x_1^3 & x_2^3 \end{bmatrix}.$$

To compute $\det(P^{(-i)}(\mathbf{x}))$, we use Cauchy-Binet formula and we have

$$\det(P^{(-i)}(\mathbf{x})) = \sum_T \det(C^{(-i)}_{:,T}) \cdot \det(X^{[\ell]}_{T,:})$$

where the summation is over all subset T of size $\ell-1$ of $[\ell]$, $C^{(i\to j)}_{:,T}$ is the $(\ell-1)$-by-$(\ell-1)$ matrix whose columns are the columns of $C^{(-i)}$ at indices from T and $X^{[\ell]}_{T,:}$ is the $(\ell-1)$-by-$(\ell-1)$ matrix whose rows are the rows of $X^{[\ell]}$ at indices from T. Furthermore, by triangle inequality,

$$|\det(P^{(-i)}(\mathbf{x}))| \le \sum_T |\det(C^{(-i)}_{:,T})| \cdot |\det(X^{[\ell]}_{T,:})| \tag{31}$$

We first make some simplifications to see what T makes the determinants nonzero. For example, take $\ell = 8, i = 2$, we have

$$
\begin{aligned}
h_0(x) &= 1 \\
h_1(x) &= x \\
h_3(x) &= -3x + x^3 \\
h_4(x) &= 3 - 6x^2 + x^4 \\
h_5(x) &= 15x - 10x^3 + x^5 \\
h_6(x) &= -15 + 45x^2 - 15x^4 + x^6 \\
h_7(x) &= -105x + 105x^3 - 21x^5 + x^7
\end{aligned}
$$

and

$$
C^{(-i)} =_{\text{up to row and column swaps}} \begin{bmatrix}
1 & & & & & & & \\
3 & -6 & 1 & & & & & \\
-15 & 45 & -15 & 1 & & & & \\
& & & & 1 & & & \\
& & & & -3 & 1 & & \\
& & & & 15 & -10 & 1 & \\
& & & & -105 & 105 & -21 & 1
\end{bmatrix}
$$

Fro simplicity we assume that i, ℓ are even numbers and it is easy to prove the other cases by symmetry. If T does not contain all odd numbers or all even numbers less than i, then $\det(C^{(-i)}_{:,T}) = 0$. In the words, the choices are $[\ell]\backslash\{b\}$ for $b = i, i+2, \ldots, \ell - 2$. Therefore, there are only $\frac{\ell - i}{2} = O(\ell)$ choices for T such that $\det(C^{(-i)}_{:,T})$ may be be 0.

Now, we expand the determinant $\det(C^{(-i)}_{:,T})$ along the rows whose diagonal entry is 1. What we have left is the determinant of a matrix A where is A is the $(\frac{b-i}{2})$-by-$(\frac{b-i}{2})$ matrix whose (r, c)-entry is $(-1)^{\frac{r-c}{2}} \frac{r!}{(\frac{r-c}{2})!c!2^{\frac{r-c}{2}}}$ for $r = i+2, \ldots, b$ and $c = i, i+2, \ldots, b-2$. For example, take $b = 6$, the matrix A in the above example is $\begin{bmatrix} -6 & 1 \\ 45 & -15 \end{bmatrix}$. By applying row and column operations, we can compute the exact expression for $\det(A)$ as

$$
\det(A) = (-1)^{\frac{b-i}{2}} \frac{b!}{(\frac{b-i}{2})!i!2^{\frac{b-i}{2}}}
$$

and hence

$$
|\det(C^{(-i)}_{:,T})| = |\det(A)| \le \frac{b!}{(\frac{b-i}{2})!i!2^{\frac{b-i}{2}}} \le \ell!. \tag{32}
$$

In the example, we have

$$
\det(\begin{bmatrix} -6 & 1 \\ 45 & -15 \end{bmatrix}) = 45.
$$

By the properties of Schur polynomials,

$$
\det(X^{[\ell]}_{T,:}) = \left(\sum_Y \mathbf{x}^Y\right) \cdot \prod_{1 \le c_1 < c_2 \le \ell - 2} (x_{c_1} - x_{c_2})
$$

where the summation is over all semi-standard Young tableaux Y of shape $(\underbrace{1, \ldots, 1}_{\ell - 1 - b \text{ 1's}}, \underbrace{0, \ldots, 0}_{b \text{ 0's}})$. Recall that the term $\mathbf{x}^Y$ means $x_0^{y_0} \cdots x_{\ell-2}^{y_{\ell-2}}$ where y_m is the number of occurrences of the number m in Y and note that $\sum_{m=0}^{\ell-2} y_m = \ell - 1 - b$.

Based on the given shape, there is only one column of size $\ell - 1 - b$. That means the number of semi-standard Young tableaux of such shape is the number of increasing sequences of size $\ell - 1 - b$ whose numbers are between 0 and $\ell - 2$ inclusive which is $\binom{\ell-1}{b} \leq 2^{\ell}$. By the assumption that $S = [-R, R]$, we can also bound the term $|\mathbf{x}^Y|$ to be

$$|\mathbf{x}^Y| \leq R^{\ell-1-b} \leq 2^{O(\ell)}.$$

It means that

$$|\det(X^{[\ell]}_{T,:})| \leq 2^{O(\ell)} \cdot \prod_{1 \leq c_1 < c_2 \leq \ell - 2} (x_{c_1} - x_{c_2}). \tag{33}$$

By plugging (32) and (33) into (31), we can now bound $|\det(P^{(-i)}(\mathbf{x}))|$ by

$$|\det(P^{(-i)}(\mathbf{x}))| \leq \sum_{T} |\det(C^{(-i)}_{:,T})| \cdot |\det(X^{[\ell]}_{T,:})| \leq 2^{O(\ell \log \ell)} \cdot \prod_{1 \leq c_1 < c_2 \leq \ell - 2} |x_{c_1} - x_{c_2}|.$$

□