Learning Mixtures of Gaussians in High Dimensions

Rong Ge[*]   Qingqing Huang[†]   Sham M. Kakade[‡]

Abstract

Efficiently learning mixture of Gaussians is a fundamental problem in statistics and learning theory. Given samples coming from a random one out of k Gaussian distributions in $\mathbb{R}^n$, the learning problem asks to estimate the means and the covariance matrices of these Gaussians. This learning problem arises in many areas ranging from the natural sciences to the social sciences, and has also found many machine learning applications.

Unfortunately, learning mixture of Gaussians is an information theoretically hard problem: in order to learn the parameters up to a reasonable accuracy, the number of samples required is exponential in the number of Gaussian components in the worst case. In this work, we show that provided we are in high enough dimensions, the class of Gaussian mixtures is learnable in its most general form under a smoothed analysis framework, where the parameters are randomly perturbed from an adversarial starting point.

In particular, given samples from a mixture of Gaussians with randomly perturbed parameters, when $n \geq \Omega(k^2)$, we give an algorithm that learns the parameters with polynomial running time and using polynomial number of samples.

The central algorithmic ideas consist of new ways to decompose the moment tensor of the Gaussian mixture by exploiting its structural properties. The symmetries of this tensor are derived from the combinatorial structure of higher order moments of Gaussian distributions (sometimes referred to as Isserlis' theorem or Wick's theorem). We also develop new tools for bounding smallest singular values of structured random matrices, which could be useful in other smoothed analysis settings.

---

[*]Microsoft Research, New England. Email: rongge@microsoft.com

[†]MIT EECS. Email:qqh@mit.edu. Part of this work was done when the author was interning at Microsoft Research New England

[‡]Microsoft Research, New England. Email: skakade@microsoft.com

1   Introduction

Learning mixtures of Gaussians is a fundamental problem in statistics and learning theory, whose study dates back to Pearson (1894). Gaussian mixture models arise in numerous areas including physics, biology and the social sciences (McLachlan and Peel (2004); Titterington et al. (1985)), as well as in image processing (Reynolds and Rose (1995)) and speech (Permuter et al. (2003)).

In a Gaussian mixture model, there are k unknown n-dimensional multivariate Gaussian distributions. Samples are generated by first picking one of the k Gaussians, then drawing a sample from that Gaussian distribution. Given samples from the mixture distribution, our goal is to estimate the means and covariance matrices of these underlying Gaussian distributions[1].

This problem has a long history in theoretical computer science. The seminal work of Dasgupta (1999) gave an algorithm for learning spherical Gaussian mixtures when the means are well separated. Subsequent works (Dasgupta and Schulman (2000); Sanjeev and Kannan (2001); Vempala and Wang (2004); Brubaker and Vempala (2008)) developed better algorithms in the well-separated case, relaxing the spherical assumption and the amount of separation required.

When the means of the Gaussians are not separated, after several works (Belkin and Sinha (2009); Kalai et al. (2010)), Belkin and Sinha (2010) and Moitra and Valiant (2010) independently gave algorithms that run in polynomial time and with polynomial number of samples for a fixed number of Gaussians. However, both running time and sample complexity depend *super* exponentially on the number of components k[2]. Their algorithm is based on the *method of moments* introduced by Pearson (1894): first estimate the $O(k)$-order moments of the distribution, then try to find the parameters that agree with these moments. Moitra and Valiant (2010) also show that the exponential dependency of the sample complexity on the number of components is necessary, by constructing an example of two mixtures of Gaussians with very different parameters, yet with exponentially small statistical distance.

Recently, Hsu and Kakade (2013) applied spectral methods to learning mixture of spherical Gaussians. When $n \geq k + 1$ and the means of the Gaussians are linearly independent, their algorithm can learn the model in polynomial time and with polynomial number of samples. This result suggests that the lower bound example in Moitra and Valiant (2010) is only a *degenerate* case in high dimensional space. In fact, *most* (in general position) mixture of spherical Gaussians are *easy* to learn. This result is also based on the method of moments, and only uses second and third moments. Several follow-up works (Bhaskara et al. (2014); Anderson et al. (2013)) use higher order moments to get better dependencies on n and k.

However, the algorithm in Hsu and Kakade (2013) as well as in the follow-ups all make strong requirements on the covariance matrices. In particular, most of them only apply to learning mixture of spherical Gaussians. For mixture of Gaussians with general covariance matrices, the best known result is still Belkin and Sinha (2010) and Moitra and Valiant (2010), which algorithms are not polynomial in the number of components k. This leads to the following natural question:

Question: *Is it possible to learn* most *mixture of Gaussians in polynomial time using a polynomial number of samples?*

Our Results   In this paper, we give an algorithm that learns *most* mixture of Gaussians in high dimensional space (when $n \geq \Omega(k^2)$), and the argument is formalized under the *smoothed analysis* framework first proposed in Spielman and Teng (2004).

---

[1] This is different from the problem of *density estimation* considered in Feldman et al. (2006); Chan et al. (2014)

[2] In fact, it is in the order of $O(e^{O(k)^k})$ as shown in Theorem 11.3 in Valiant (2012).

In the smoothed analysis framework, the adversary first choose an arbitrary mixture of Gaussians. Then the mean vectors and covariance matrices of this Gaussian mixture are randomly *perturbed* by a small amount ρ [3]. The samples are then generated from the Gaussian mixture model with the perturbed parameters. The goal of the algorithm is to learn the perturbed parameters from the samples.

The smoothed analysis framework is a natural bridge between worst-case and average-case analysis. On one hand, it is similar to worst-case analysis, as the adversary chooses the initial instance, and the perturbation allowed is small. On the other hand, even with small perturbation, we may hope that the instance be different enough from degenerate cases. A successful algorithm in the smoothed analysis setting suggests that the bad instances must be very "sparse" in the parameter space: they are highly unlikely in any small neighborhood of any instance. Recently, the smoothed analysis framework has also motivated several research work (Kalai et al. (2009) Bhaskara et al. (2014)) in analyzing learning algorithms.

In the smoothed analysis setting, we show that it is easy to learn most Gaussian mixtures:

Theorem 1.1. *(informal statement of Theorem 3.4) In the smoothed analysis setting, when $n \geq \Omega(k^2)$, given samples from the perturbed n-dimensional Gaussian mixture model with k components, there is an algorithm that learns the correct parameters up to accuracy ϵ with high probability, using polynomial time and number of samples.*

An important step in our algorithm is to learn Gaussian mixture models whose components all have mean zero, which is also a problem of independent interest (Zoran and Weiss (2012)). Intuitively this is also a "hard" case, as there is no separation in the means. Yet algebraically, this case gives rise to a novel tensor decomposition algorithm. The ideas for solving this decomposition problem are then generalized to tackle the most general case.

Theorem 1.2. *(informal statement of Theorem 3.5) In the smoothed analysis setting, when $n \geq \Omega(k^2)$, given samples from the perturbed mixture of zero-mean n-dimensional Gaussian mixture model with k components, there is an algorithm that learns the parameters up to accuracy ϵ with high probability, using polynomial running time and number of samples.*

Organization   The main part of the paper will focus on learning mixtures of zero-mean Gaussians. The proposed algorithm for this special case contains most of the new ideas and techniques. In Section 2 we introduce the notations for matrices and tensors which are used to handle higher order moments throughout the discussion. Then in Section 3 we introduce the smoothed analysis model for learning mixture of Gaussians and discuss the moment structure of mixture of Gaussians, then we formally state our main theorems. Section 4 outlines our algorithm for learning zero-mean mixture of Gaussians. The details of the steps are presented in Section 5. The detailed proofs for the correctness and the robustness are deferred to Appendix (Sections B to D). In Section 6 we briefly discuss how the ideas for zero-mean case can be generalized to learning mixture of nonzero Gaussians, for which the detailed algorithm and the proofs are deferred to Appendix F.

2   Notations

Vectors and Matrices   In the vector space $\mathbb{R}^n$, let $\langle \cdot, \cdot \rangle$ denote the inner product of two vectors, and $\| \cdot \|$ to denote the Euclidean norm.

---

[3]See Definition 3.2 in Section 3.1 for the details.

For a tall matrix $A \in \mathbb{R}^{m \times n}$, let $A_{[:,j]}$ denote its j-th column vector, let $A^\top$ denote its transpose, $A^\dagger = (A^\top A)^{-1} A^\top$ denote the pseudoinverse, and let $\sigma_k(A)$ denote its k-th singular value. Let I_n be the identity matrix of dimension $n \times n$. The spectral norm of a matrix is denoted as $\| \cdot \|$, and the Frobenius norm is denoted as $\| \cdot \|_F$. We use $A \succeq 0$ for positive semidefinite matrix A.

In the discussion, we often need to convert between vectors and matrices. Let $\text{vec}(A) \in \mathbb{R}^{mn}$ denote the vector obtained by stacking all the columns of A. For a vector $x \in \mathbb{R}^{m^2}$, let $\text{mat}(x) \in \mathbb{R}^{m \times m}$ denote the inverse mapping such that $\text{vec}(\text{mat}(x)) = x$.

We use $[n]$ to denote the set $\{1, 2, ..., n\}$ and $[n] \times [n]$ to denote the set $\{(i, j) : i, j \in [n]\}$. These are often used as indices of matrices.

Symmetric matrices   We use $\mathbb{R}^{n \times n}_{sym}$ to denote the space of all $n \times n$ symmetric matrices, which subspace has dimension $\binom{n+1}{2}$. Since we will frequently use $n \times n$ and $k \times k$ symmetric matrices, we denote their dimensions by the constants $n_2 = \binom{n+1}{2}$ and $k_2 = \binom{k+1}{2}$. Similarly, we use $\mathbb{R}^{n \times \cdots \times n}_{sym}$ to denote the symmetric k-dimensional multi-arrays (tensors), which subspace has dimension $\binom{n+k-1}{k}$. If a k-th order tensor $X \in \mathbb{R}^{n \times \cdots \times n}_{sym}$, then for any permutation π over $[k]$, we have $X_{n_1,...,n_k} = X_{n_{\pi(1)},...,n_{\pi(k)}}$.

Linear subspaces   We represent a linear subspace $\mathcal{S} \in \mathbb{R}^n$ of dimension d by a matrix $S \in \mathbb{R}^{n \times d}$, whose columns of S form an (arbitrary) orthonormal basis of the subspace. The projection matrix onto the subspace $\mathcal{S}$ is denoted by $\text{Proj}_S = SS^\top$, and the projection onto the orthogonal subspace $\mathcal{S}^\perp$ is denoted by $\text{Proj}_{S^\perp} = I_n - SS^\top$. When we talk about the span of several matrices, we mean the space spanned by their vectorization.

Tensors   A tensor is a multi-dimensional array. Tensor notations are useful for handling higher order moments. We use $\otimes$ to denote tensor product, suppose $a, b, c \in \mathbb{R}^n$, $T = a \otimes b \otimes c \in \mathbb{R}^{n \times n \times n}$ and $T_{i_1, i_2, i_3} = a_{i_1} b_{i_2} c_{i_3}$. For a vector $x \in \mathbb{R}^n$, let the t-fold tensor product $x \otimes^t$ denote the t-th order rank one tensor $(x \otimes^t)_{i_1, i_2, ..., i_t} = \prod_{j=1}^t x_{i_j}$.

Every tensor defines a multilinear mapping. Consider a 3-rd order tensor $X \in \mathbb{R}^{n_A \times n_B \times n_C}$. For given dimension m_A, m_B, m_C, it defines a multi-linear mapping $X(\cdot, \cdot, \cdot) : \mathbb{R}^{n_A \times m_A} \times \mathbb{R}^{n_B \times m_B} \times \mathbb{R}^{n_C \times m_C} \to \mathbb{R}^{m_A \times m_B \times m_C}$ defined as below: ($\forall j_1 \in [m_A], j_2 \in [m_B], j_3 \in [m_C]$)

$$[X(V_1, V_2, V_3)]_{j_1, j_2, j_3} = \sum_{i_1 \in [n_A], i_2 \in [n_B], i_3 \in [n_C]} X_{i_1, i_2, i_3} [V_1]_{j_1, i_1} [V_2]_{j_2, i_2} [V_3]_{j_3, i_3}.$$

If X admits a decomposition $X = \sum_{i=1}^k A_{[:,i]} \otimes B_{[:,i]} \otimes C_{[:,i]}$ for $A \in \mathbb{R}^{n_A \times k}, B \in \mathbb{R}^{n_B \times k}, C \in \mathbb{R}^{n_C \times k}$, the multi-linear mapping has the form $X(V_1, V_2, V_3) = \sum_{i=1}^k (V_1^\top A_{[:,i]}) \otimes (V_2^\top B_{[:,i]}) \otimes (V_3^\top C_{[:,i]})$.

In particular, the vector given by $X(\mathbf{e}_i, \mathbf{e}_j, I)$ is the one-dimensional slice of the 3-way array, with the index for the first dimension to be i and the second dimension to be j.

Matrix Products We use $\odot$ to denote column wise Katri-Rao product, and $\otimes_{kr}$ to denote Kronecker product. As an example, for matrices $A \in \mathbb{R}^{m_A \times n}$, $B \in \mathbb{R}^{m_B \times n}$, $C \in \mathbb{R}^{m_C \times n}$:

$$A \otimes B \otimes C \in \mathbb{R}^{m_A \times m_B \times m_C}, \quad [A \otimes B \otimes C]_{j_1, j_2, j_3} = \sum_{i=1}^{n} A_{j_1, i} B_{j_2, i} C_{j_3, i},$$

$$A \odot B \in \mathbb{R}^{m_A m_B \times n}, \quad [A \odot B]_{[:,j]} = A_{[:,j]} \otimes_{kr} B_{[:,j]}.$$

$$A \otimes_{kr} B \in \mathbb{R}^{m_A m_B \times n^2}, \quad A \otimes_{kr} B = \begin{bmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m_A,1}B & \cdots & A_{m_A,n}B \end{bmatrix},$$

3 Main results

In this section, we first formally introduce the smoothed analysis framework for our problem and state our main theorems. Then we will discuss the structure of the moments of Gaussian mixtures, which is crucial for understanding our method of moments based algorithm.

3.1 Smoothed Analysis for Learning Mixture of Gaussians

Let $\mathcal{G}_{n,k}$ denote the class of Gaussian mixtures with k components in $\mathbb{R}^n$. A distribution in this family is specified by the following parameters: the mixing weights ω_i, the mean vectors $\mu^{(i)}$ and the covariance matrices $\Sigma^{(i)}$, for $i \in [k]$.

$$\mathcal{G}_{n,k} := \left\{ \mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)})\}_{i \in [k]} : \omega_i \in \mathbb{R}_+, \ \sum_{i=1}^{k} \omega_i = 1, \ \mu^{(i)} \in \mathbb{R}^n, \ \Sigma^{(i)} \in \mathbb{R}_{sym}^{n \times n}, \ \Sigma^{(i)} \succeq 0 \right\}.$$

As an interesting special case of the general model, we also consider the mixture of "zero-mean" Gaussians, which has $\mu^{(i)} = 0$ for all components $i \in [k]$.

A sample x from a mixture of Gaussians is generated in two steps:

1. Sample $h \in [k]$ from a multinomial distribution, with probability $\Pr[h = i] = \omega_i$ for $i \in [k]$.

2. Sample $x \in \mathbb{R}^n$ from the h-th Gaussian distribution $\mathcal{N}(\mu^{(h)}, \Sigma^{(h)})$.

The learning problem asks to estimate the parameters of the underlying mixture of Gaussians:

Definition 3.1 (Learning mixture of Gaussians). *Given N samples $x_1, x_2, ..., x_N$ drawn i.i.d. from a mixture of Gaussians $\mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)})\}_{i \in [k]}$, an algorithm learns the mixture of Gaussians with accuracy ϵ, if it outputs an estimation $\widehat{\mathcal{G}} = \{(\widehat{\omega}_i, \widehat{\mu}^{(i)}, \widehat{\Sigma}^{(i)})\}_{i \in [k]}$ such that there exists a permutation π on $[k]$, and for all $i \in [k]$, we have $|\widehat{\omega}_i - \omega_{\pi(i)}| \leq \epsilon$, $\|\widehat{\mu}^{(i)} - \mu^{(\pi(i))}\| \leq \epsilon$ and $\|\widehat{\Sigma}^{(i)} - \Sigma^{(\pi(i))}\| \leq \epsilon$.*

In the worst case, learning mixture of Gaussians is a information theoretically hard problem (Moitra and Valiant (2010)). There exists worst-case examples where the number of samples required for learning the instance is at least exponential in the number of components k (McLachlan and Peel (2004)). The non-convexity arises from the hidden variable h: without knowing h we cannot determine which Gaussian component each sample comes from.

The smoothed analysis framework provides a way to circumvent the worst case instances, yet still studying this problem in its most general form. The basic idea is that, with high probability over the small random perturbation to any instance, the instance will not be a "worst-case" instance, and actually has reasonably good condition for the algorithm.

Next, we show how the parameters of the mixture of Gaussians are *perturbed* in our setup.

Definition 3.2 (ρ-smooth mixture of Gaussian). *For $\rho < 1/n$, a ρ-smooth n-dimensional k-component mixture of Gaussians $\widetilde{\mathcal{G}} = \{(\widetilde{\omega}_i, \widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$ is generated as follows:*

1. *Choose an arbitrary (could be adversarial) instance $\mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$. Scale the distribution such that $0 \preceq \Sigma^{(i)} \preceq \frac{1}{2} I_n$ and $\|\mu^{(i)}\| \leq \frac{1}{2}$ for all $i \in [k]$.*

2. *Let $\Delta_i \in \mathbb{R}^{n \times n}_{sym}$ be a random symmetric matrix with zeros on the diagonals, and the upper-triangular entries are independent random Gaussian variables $\mathcal{N}(0, \rho^2)$. Let $\delta_i \in \mathbb{R}^n$ be a random Gaussian vector with independent Gaussian variables $\mathcal{N}(0, \rho^2)$.*

3. *Set $\widetilde{\omega}_i = \omega_i$, $\widetilde{\mu}^{(i)} = \mu^{(i)} + \delta_i$, $\widetilde{\Sigma}^{(i)} = \Sigma^{(i)} + \Delta_i$.*

4. *Choose the diagonal entries of $\widetilde{\Sigma}^{(i)}$ arbitrarily, while ensuring the positive semi-definiteness of the covariance matrix $\widetilde{\Sigma}^{(i)}$, and the diagonal entries are upper bounded by 1. The perturbation procedure fails if this step is infeasible[4].*

A ρ-smooth zero-mean mixture of Gaussians is generated using the same procedure, except that we set $\widetilde{\mu}^{(i)} = \mu^{(i)} = 0$, for all $i \in [k]$.

Remark 3.3. *When the original matrix is of low rank, a simple random perturbation may not lead to a positive semidefinite matrix, which is why our procedure of perturbation is more restricted in order to guarantee that the perturbed matrix is still a valid covariance matrix.*

There could be other ways of locally perturbing the covariance matrix. Our procedure actually gives more power to the adversary as it can change the diagonals after observing the perturbations for other entries. Note that with high probability if we just let the new diagonal to be $5\sqrt{n}\rho$ larger than the original ones, the resulting matrix is still a valid covariance matrix. In other words, the adversary can always keep the perturbation small if it wants to.

Instead of the worst-case problem in Definition 3.1, our algorithms work on the smoothed instance. Here the model first gets perturbed to $\widetilde{\mathcal{G}} = \{(\widetilde{\omega}_i, \widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)})\}_{i \in [k]}$, the samples are drawn according to the perturbed model, and the algorithm tries to learn the perturbed parameters. We give a polynomial time algorithm in this case:

Theorem 3.4 (Main theorem). *Consider a ρ-smooth mixture of Gaussians $\widetilde{\mathcal{G}} = \{(\widetilde{\omega}_i, \widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$ for which the number of components is at least [5] $k \geq C_0$ and the dimension $n \geq C_1 k^2$, for some fixed constants C_0 and C_1. Suppose that the mixing weights $\widetilde{\omega}_i \geq \omega_o$ for all $i \in [k]$. Given N samples drawn i.i.d. from $\widetilde{\mathcal{G}}$, there is an algorithm that learns the parameters of $\widetilde{\mathcal{G}}$ up to accuracy ϵ, with high probability over the randomness in both the perturbation and the samples. Furthermore, the running time and number of samples N required are both upper bounded by $poly(n, k, 1/\omega_o, 1/\epsilon, 1/\rho)$.*

To better illustrate the algorithmic ideas for the general case, we first present an algorithm for learning mixtures of zero-mean Gaussians. Note that this is not just a special case of the general case, as with the smoothed analysis, the zero mean vectors are not perturbed.

Theorem 3.5 (Zero-mean). *Consider a ρ-smooth mixture of zero-mean Gaussians $\widetilde{\mathcal{G}} = \{(\widetilde{\omega}_i, 0, \widetilde{\Sigma}^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$ for which the number of components is at least $k \geq C_0$ and the dimension $n \geq C_1 k^2$, for some fixed constants C_0 and C_1. Suppose that the mixing weights $\widetilde{\omega}_i \geq \omega_o$ for all $i \in [k]$. Given N*

---

[4] Note that by standard random matrix theory, with high probability the 4-th step is feasible and the perturbation procedure in Definition 3.2 succeeds. Also, with high probability we have $\|\widetilde{\mu}^{(i)}\| \leq 1$ and $0 \preceq \widetilde{\Sigma}^{(i)} \preceq I_n$ for all $i \in [k]$.

[5] Note that the algorithms of Belkin and Sinha (2010) and Moitra and Valiant (2010) run in polynomial time for fixed k.

samples drawn i.i.d. from $\widetilde{\mathcal{G}}$, there is an algorithm that learns the parameters of $\widetilde{\mathcal{G}}$ up to accuracy ϵ, with high probability over the randomness in both the perturbation and the samples. Furthermore, the running time and number of samples N are both upper bounded by $poly(n, k, 1/\omega_o, 1/\epsilon, 1/\rho)$.

Throughout the paper we always assume that $n \geq C_1 k^2$ and $\widetilde{\omega}_i \geq \omega_o$.

3.2 Moment Structure of Mixture of Gaussians

Our algorithm is also based on the method of moments, and we only need to estimate the 3-rd, the 4-th and the 6-th order moments. In this part we briefly discuss the structure of 4-th and 6-th moments in the zero-mean case (3-rd moment is always 0 in the zero-mean case). These structures are essential to the proposed algorithm. For more details, and discussions on the general case see Appendix A.

The m-th order moments of the *zero-mean* Gaussian mixture model $\mathcal{G} \in \mathcal{G}_{n,k}$ are given by the following m-th order symmetric tensor $M_m \in \mathbb{R}_{sym}^{n \times \cdots \times n}$:

$$[M_m]_{j_1, \ldots, j_m} := \mathbb{E}[x_{j_1} \ldots x_{j_m}] = \sum_{i=1}^{k} \omega_i \mathbb{E}\left[y_{j_1}^{(i)} \ldots y_{j_m}^{(i)}\right], \quad \forall j_1, \ldots, j_m \in [n],$$

where $y^{(i)}$ corresponds to the n-dimensional zero-mean Gaussian distribution $\mathcal{N}(0, \Sigma^{(i)})$. The moments for each Gaussian component are characterized by Isserlis's theorem as below:

Theorem 3.6 (Isserlis' Theorem). *Let $(y_1, \ldots, y_{2t})$ be a multivariate zero-mean Gaussian random vector $\mathcal{N}(0, \Sigma)$, then*

$$\mathbb{E}[y_1 \ldots y_{2t}] = \sum \prod \Sigma_{u,v},$$

where the summation is taken over all distinct ways of partitioning $y_1, \ldots, y_{2t}$ into t pairs, which correspond to all the perfect matchings in a complete graph.

Ideally, we would like to obtain the following quantities (recall $n_2 = \binom{n+1}{2}$):

$$X_4 = \sum_{i=1}^{k} \omega_i \mathrm{vec}(\Sigma^{(i)})^{\otimes 2} \in \mathbb{R}^{n_2 \times n_2}, \quad X_6 = \sum_{i=1}^{k} \omega_i \mathrm{vec}(\Sigma^{(i)})^{\otimes 3} \in \mathbb{R}^{n_2 \times n_2 \times n_2}. \tag{1}$$

Note that the entries in X_4 and X_6 are quadratic and cubic monomials of the covariance matrices, respectively. If we have X_4 and X_6, the tensor decomposition algorithm in Anandkumar et al. (2014) can be immediately applied to recover ω_i's and $\Sigma^{(i)}$'s under mild conditions. It is easy to verify that those conditions are indeed satisfied with high probability in the smoothed analysis setting.

By Isserlis's theorem, the entries of the moments M_4 and M_6 are indeed quadratic and cubic functions of the covariance matrices, respectively. However, the structure of the true moments M_4 and M_6 have more symmetries, consider for example,

$$[M_4]_{1,2,3,4} = \sum_{i=1}^{k} \omega_i (\Sigma_{1,2}^{(i)} \Sigma_{3,4}^{(i)} + \Sigma_{1,3}^{(i)} \Sigma_{2,4}^{(i)} + \Sigma_{1,4}^{(i)} \Sigma_{2,3}^{(i)}), \quad \text{while } [X_4]_{(1,2),(3,4)} = \sum_{i=1}^{k} \omega_i \Sigma_{1,2}^{(i)} \Sigma_{3,4}^{(i)}.$$

Note that due to symmetry, the number of distinct entries in M_4 ($\binom{n+3}{4} \approx n^4/24$) is much smaller than the number of distinct entries in X_4 ($\binom{n_2+1}{2} \approx n^4/8$). Similar observation can be made about M_6 and X_6.

Therefore, it is not immediate how to find the desired X_4 and X_6 based on M_4 and M_6. We call the moments M_4, M_6 the *folded moments* as they have more symmetry, and the corresponding X_4, X_6 the *unfolded moments*. One of the key steps in our algorithm is to unfold the true moments M_4, M_6 to get X_4, X_6 by exploiting special structure of M_4, M_6.

In some cases, it is easier to restrict our attention to the entries in M_4 with indices corresponding to distinct variables. In particular, we define

$$\overline{M}_4 = [[M_4]_{j_1,j_2,j_3,j_4} : 1 \leq j_1 < j_2 < j_3 < j_4 \leq n] \in \mathbb{R}^{n_4}, \tag{2}$$

where $n_4 = \binom{n}{4}$ is the number of 4-tuples with indices corresponding to distinct variables. We define $\overline{M}_6 \in \mathbb{R}^{n_6}$ similarly where $n_6 = \binom{n}{6}$. We will see that these entries are nice as they are *linear projections* of the desired unfolded moments X_4 and X_6 (Lemma 3.7 below), also such projections satisfy certain "symmetric off-diagonal" properties which are convenient for the proof (see Definition C.3 in Section C).

Lemma 3.7. *For a zero-mean Gaussian mixture model, there exist two fixed and known linear mappings $\mathcal{F}_4 : \mathbb{R}^{n_2 \times n_2} \to \mathbb{R}^{n_4}$ and $\mathcal{F}_6 : \mathbb{R}^{n_2 \times n_2 \times n_2} \to \mathbb{R}^{n_6}$ such that:*

$$\overline{M}_4 = \sqrt{3}\mathcal{F}_4(X_4), \quad \overline{M}_6 = \sqrt{15}\mathcal{F}_6(X_6). \tag{3}$$

Moreover $\mathcal{F}_4$ is a projection from a $\binom{n_2+1}{2}$-dimensional subspace to a n_4-dimensional subspace, and $\mathcal{F}_6$ is a projection from a $\binom{n_2+2}{3}$-dimensional subspace to a n_6-dimensional subspace.

4 Algorithm Outline for Learning Mixture of Zero-Mean Gaussians

In this section, we present our algorithm for learning zero-mean Gaussian mixture model. The algorithmic ideas and the analysis are at the core of this paper. Later we show that it is relatively easy to generalize the basic ideas and the techniques to handle the general case.

For simplicity we state our algorithm using the exact moments $\widetilde{M}_4$ and $\widetilde{M}_6$, while in implementation the empirical moments $\widehat{M}_4$ and $\widehat{M}_6$ obtained with the samples are used. In later sections, we verify the correctness of the algorithm and show that it is robust: the algorithm learns the parameters up to arbitrary accuracy using polynomial number of samples.

Step 1. *Span Finding: Find the span of covariance matrices .*

(a) *For a set of indices $\mathcal{H} \subset [n]$ of size $|\mathcal{H}| = \sqrt{n}$, find the span:*

$$\mathcal{S} = span\left\{\widetilde{\Sigma}^{(i)}_{[:,j]} : i \in [k], j \in \mathcal{H}\right\} \subset \mathbb{R}^n. \tag{4}$$

(b) *Find the span of the covariance matrices with the columns projected onto $\mathcal{S}^\perp$, namely,*

$$\mathcal{U}_S = span\left\{vec(Proj_{S^\perp}\widetilde{\Sigma}^{(i)}) : i \in [k]\right\} \subset \mathbb{R}^{n^2}. \tag{5}$$

(c) *For two disjoint sets of indices $\mathcal{H}_1$ and $\mathcal{H}_2$, repeat Step 1 (a) and Step 1 (b) to obtain $\mathcal{U}_1$ and $\mathcal{U}_2$, namely the span of covariance matrices projected onto two subspaces $\mathcal{S}_1^\perp$ and $\mathcal{S}_2^\perp$. Merge $\mathcal{U}_1$ and $\mathcal{U}_2$ to obtain the span of covariance matrices $\mathcal{U}$:*

$$\mathcal{U} = span\left\{\widetilde{\Sigma}^{(i)} : i \in [k]\right\} \subset \mathbb{R}^{n_2}. \tag{6}$$

7

Step 2. *Unfolding: Recover the unfolded moments $\widetilde{X}_4, \widetilde{X}_6$.*

Given the folded moments $\overline{\widetilde{M}_4}, \overline{\widetilde{M}_6}$ as defined in (2), and given the subspace $U \in \mathbb{R}^{n_2 \times k}$ from Step 1, let $\widetilde{Y}_4 \in \mathbb{R}^{k \times k}_{sym}$ and $\widetilde{Y}_6 \in \mathbb{R}^{k \times k \times k}_{sym}$ be the unknowns, solve the following systems of linear equations.

$$\overline{\widetilde{M}_4} = \sqrt{3}\mathcal{F}_4(U\widetilde{Y}_4 U^\top), \quad \overline{\widetilde{M}_6} = \sqrt{15}\mathcal{F}_6(\widetilde{Y}_6(U^\top, U^\top, U^\top)). \tag{7}$$

The unfolded moments $\widetilde{X}_4, \widetilde{X}_6$ are then given by $\widetilde{X}_4 = U\widetilde{Y}_4 U^\top, \widetilde{X}_6 = \widetilde{Y}_6(U^\top, U^\top, U^\top)$.

Step 3. *Tensor Decomposition: learn $\widetilde{\omega}_i$ and $\widetilde{\Sigma}^{(i)}$ from $\widetilde{Y}_4$ and $\widetilde{Y}_6$.*

Given U, and given $\widetilde{Y}_4$ and $\widetilde{Y}_6$ which are relate to the parameters as follows:

$$\widetilde{Y}_4 = \sum_{i=1}^{k} \widetilde{\omega}_i (U^\top \widetilde{\Sigma}^{(i)})^{\otimes 2}, \quad \widetilde{Y}_6 = \sum_{i=1}^{k} \widetilde{\omega}_i (U^\top \widetilde{\Sigma}^{(i)})^{\otimes 3},$$

we apply tensor decomposition techniques to recover $\widetilde{\Sigma}^{(i)}$'s and $\widetilde{\omega}_i$'s.

5 Implementing the Steps for Mixture of Zero-Mean Gaussians

In this part we show how to accomplish each step of the algorithm outlined in Section 4 and sketch the proof ideas.

For each step, we first explain the detailed algorithm, and list the deterministic conditions on the underlying parameters as well as on the *exact* moments for the step to work correctly. Then we show that these deterministic conditions are satisfied with high probability over the ρ-perturbation of the parameters in the smoothed analysis setting. In order to analyze the sample complexity, we further show that when we are given the *empirical* moments which are close to the exact moments, the output of the step is also close to that in the exact case.

In particular we show the correctness and the stability of each step in the algorithm with two main lemmas: the first lemma shows that with high probability over the random perturbation of the covariance matrices, the exact moments satisfy the deterministic conditions that ensure the correctness of each step; the second lemma shows that when the algorithm for each step works correctly, it is actually stable even when the moments are estimated from finite samples and have only inverse polynomial accuracy to the exact moments.

The detailed proofs are deferred to Section B to D in the appendix.

Step 1: Span Finding. Given the 4-th order moments $\widetilde{M}_4$, Step 1 finds the span of covariance matrices $\mathcal{U}$ as defined in (6). Note that by definition of the unfolded moments $\widetilde{X}_4$ in (1), the subspace $\mathcal{U}$ coincides with the column span of the matrix $\widetilde{X}_4$.

By Lemma 3.7, we know that the entries in $\widetilde{M}_4$ are linear mappings of entries in $\widetilde{X}_4$. Since the matrix $\widetilde{X}_4$ is of low rank ($k \ll n_2$), this corresponds to the *matrix sensing* problem first studied in Recht et al. (2010). In general, matrix sensing problems can be hard even when we have many linear observations (Hardt et al. (2014b)). Previous works (Recht et al. (2010); Hardt et al. (2014a); Jain et al. (2013)) showed that if the linear mapping satisfy *matrix RIP* property, one can uniquely recover $\widetilde{X}_4$ from $\widetilde{M}_4$.

However, properties like RIP do not hold in our setting where the linear mapping is determined by Isserlis' Theorem. We can construct two different mixtures of Gaussians with different unfolded moments $\widetilde{X}_4$, but the same folded moment $\widetilde{M}_4$ (see Section A.3). Therefore the existing matrix recovery algorithm cannot be applied, and we need to develop new tools by exploiting the special moment structure of Gaussian mixtures.

Step 1 (a). Find the Span of a Subset of Columns of the Covariance Matrices. The key observation for this step is that if we hit $\widetilde{M}_4$ with three basis vectors, we get a vector that lies in the span of the columns of the covariance matrices:

Claim 5.1. *For a mixture of zero-mean Gaussians $\mathcal{G} = \{(\omega_i, 0, \Sigma^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$, the one-dimensional slices of the 4-th order moments M_4 are given by:*

$$M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, \mathbf{e}_{j_3}, I) = \sum_{i=1}^{k} \omega_i \left(\Sigma_{j_1,j_2}^{(i)} \Sigma_{[:,j_3]}^{(i)} + \Sigma_{j_1,j_3}^{(i)} \Sigma_{[:,j_2]}^{(i)} + \Sigma_{j_2,j_3}^{(i)} \Sigma_{[:,j_1]}^{(i)} \right), \quad \forall j_1, j_2, j_3 \in [n]. \quad (8)$$

In particular, if we pick the indices j_1, j_2, j_3 in the index set $\mathcal{H}$, the vector $M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, \mathbf{e}_{j_3}, I)$ lies in the desired span $\mathcal{S} = \left\{ \Sigma_{[:,j]}^{(i)} : i \in [k], j \in \mathcal{H} \right\}$.

We shall partition the set $\mathcal{H}$ into three disjoint subsets $\mathcal{H}^{(i)}$ of equal size $\sqrt{n}/3$, and pick $j_i \in H^{(i)}$ for $i = 1, 2, 3$. In this way, we have $(|\mathcal{H}|/3)^3 = \Omega(n^{1.5})$ such one-dimensional slices of M_4, which all lie in the desired subspace $\mathcal{S}$. Moreover, the dimension of the subspace $\mathcal{S}$ is at most $k|\mathcal{H}| \ll n^{1.5}$. Therefore, with the ρ-perturbed parameters $\widetilde{\Sigma}^{(i)}$'s, we can expect that with high probability the slices of $\widetilde{M}_4$ span the entire subspace $\mathcal{S}$.

Condition 5.2 (Deterministic condition for Step 1 (a)). *Let $\widetilde{Q}_S \in \mathbb{R}^{n \times (|\mathcal{H}|/3)^3}$ be the matrix whose columns are the vectors $\widetilde{M}_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, \mathbf{e}_{j_3}, I)$ for $j_i \in \mathcal{H}^{(i)}$. If the matrix $\widetilde{Q}_S$ achieves its maximal column rank $k|\mathcal{H}|$, we can find the desired span $\mathcal{S}$ defined in (4) by the column span of matrix $\widetilde{Q}_S$.*

We first show that this deterministic condition is satisfied with high probability by bounding the $k|\mathcal{H}|$-th singular value of $\widetilde{Q}_S$ with smoothed analysis.

Lemma 5.3 (Correctness). *Given the exact 4-th order moments $\widetilde{M}_4$, for any index set $\mathcal{H}$ of size $|\mathcal{H}| = \sqrt{n}$, With high probability, the $k|\mathcal{H}|$-th singular value of $\widetilde{Q}_S$ is at least $\Omega(\omega_o \rho^2 n)$.*

The proof idea involves writing the matrix $\widetilde{Q}_S$ as a product of three matrices, and using the results on spectral properties of random matrices Rudelson and Vershynin (2009) to show that with high probability the smallest singular value of each factor is lower bounded.

Since this step only involves the singular value decomposition of the matrix $\widetilde{Q}_S$, we then use the standard matrix perturbation theory to show that this step is stable:

Lemma 5.4 (Stability). *Given the empirical estimator of the 4-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$, suppose that the entries of E_4 have absolute value at most δ. Let the columns of matrix $\widetilde{S} \in \mathbb{R}^{n \times k|\mathcal{H}|}$ be the left singular vector of $\widetilde{Q}_S$, and let $\widehat{S}$ be the corresponding matrix obtained with $\widehat{M}_4$. When δ is inverse polynomially small, the distance between the two projections $\|Proj_{\widehat{S}} - Proj_{\widetilde{S}}\|$ is upper bounded by $O\left(n^{1.25} \delta / \sigma_{k|\mathcal{H}|}(\widetilde{Q}_S) \right)$.*

Remark 5.5. *Note that we need the high dimension assumption $(n \gg k)$ to guarantee the correctness of this step: in order to span the subspace $\mathcal{S}$, the number of distinct vectors should be equal or larger than the dimension of the subspace, namely $|\mathcal{H}|^3 \geq k|\mathcal{H}|$; and the subspace should be non-trivial, namely $k|\mathcal{H}| < n$. These two inequalities suggest that we need $n \geq \Omega(k^{1.5})$. However, we used the stronger assumption $n \geq \Omega(k^2)$ to obtain the lower bound of the smallest singular value in the proof.*

Step 1 (b). Find the Span of Projected Covariance Matrices. In this step, we continue to use the structural properties of the 4-th order moments. In particular, we look at the two-dimensional slices of M_4 obtained by hitting it with two basis vectors:

Claim 5.6. *For a mixture of zero-mean Gaussians $\mathcal{G} = \{(\omega_i, 0, \Sigma^{(i)})\}_{i \in [k]} \in \mathcal{G}_{n,k}$, the two-dimensional slices of the 4-th order moments M_4 are given by:*

$$M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I) = \sum_{i=1}^{k} \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma^{(i)} + \Sigma_{[:,j_1]}^{(i)} (\Sigma_{[:,j_2]}^{(i)})^\top + \Sigma_{[:,j_2]}^{(i)} (\Sigma_{[:,j_1]}^{(i)})^\top \right), \quad \forall j_1, j_2 \in [n]. \quad (9)$$

Note that if we take the indices j_1 and j_2 in the index set $\mathcal{H}$, the slice $M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I)$ is *almost* in the span of the covariance matrices, except $2k$ additive rank-one terms in the form of $\Sigma_{[:,j_1]}^{(i)} (\Sigma_{[:,j_2]}^{(i)})^\top$. These rank-one terms can be eliminated by projecting the slice to the subspace $\mathcal{S}^\perp$ obtained in Step 1 (a), namely,

$$\text{vec}(\text{Proj}_{S^\perp} M_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I)) = \sum_{i=1}^{k} \omega_i \Sigma_{j_1, j_2}^{(i)} \text{vec}(\text{Proj}_{S^\perp} \Sigma^{(i)}), \quad \forall j_1, j_2 \in \mathcal{H},$$

and this projected two-dimensional slice lies in the desired span $\mathcal{U}_S$ as defined in (5). Moreover, there are $\binom{|\mathcal{H}|+1}{2} = \Omega(n)$ such projected two-dimensional slices, while the dimension of the desired span $\mathcal{U}_S$ is at most k.

Condition 5.7 (Deterministic condition for Step 1 (b)). *Let $\widetilde{Q}_{U_S} \in \mathbb{R}^{n_2 \times |\mathcal{H}|(|\mathcal{H}|+1)/2}$ be a matrix whose (j_1, j_2)-th column for is equal to the projected two-dimensional slice $\text{vec}(\text{Proj}_{S^\perp} \widetilde{M}_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I))$, for $j_1 \leq j_2$ and $j_1, j_2 \in \mathcal{H}$. If the matrix $\widetilde{Q}_{U_S}$ achieves its maximal column rank k, the desired span $\mathcal{U}_S$ defined in (5) is given by the column span of the matrix $\widetilde{Q}_{U_S}$.*

We show that this deterministic condition is satisfied by bounding the k-th singular value of $\widetilde{Q}_{U_S}$ in the smoothed analysis setting:

Lemma 5.8 (Correctness). *Given the exact 4-th order moments $\widetilde{M}_4$, with high probability, the k-th singular value of $\widetilde{Q}_{U_S}$ is at least $\Omega(\omega_o \rho^2 n^{1.5})$.*

Similar to Lemma 5.3, the proof is based on writing the matrix Q_{U_S} as a product of three matrices, then bound their k-th singular values using random matrix theory. The stability analysis also relies on the matrix perturbation theory.

Lemma 5.9 (Stability). *Given the empirical 4-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$, assume that the absolute value of entries of E_4 are at most δ_2. Also, given the output $\text{Proj}_{\widehat{S}^\perp}$ from Step 1 (a), and assume that $\|\text{Proj}_{\widehat{S}^\perp} - \text{Proj}_{\widetilde{S}^\perp}\| \leq \delta_1$. When δ_1 and δ_2 are inverse polynomially small, we have $\|\text{Proj}_{\widehat{U}_S} - \text{Proj}_{\widetilde{U}_S}\| \leq O\left(n^{2.5} (\delta_2 + 2\delta_1) / \sigma_k(\widetilde{Q}_{U_S})\right)$.*

Step 1 (c). Merge $\mathcal{U}_1, \mathcal{U}_2$ to get the span of covariance matrices $\mathcal{U}$. Note that for a given index set $\mathcal{H}$, the span $\mathcal{U}_S$ obtained in Step 1 (b) only gives partial information about the span of the covariance matrices. The idea of getting the span of the full covariance matrices is to obtain two sets of such partial information and then merge them.

In order to achieve that, we repeat Step 1 (a) and Step 1 (b) for two *disjoint* sets $\mathcal{H}_1$ and $\mathcal{H}_2$, each of size $\sqrt{n}$. The two subspace S_1 and S_2 thus correspond to the span of two disjoint sets of covariance matrix columns. Therefore, we can hope that U_1 and U_2, the span of covariance matrices projected to $S_1^\perp$ and $S_2^\perp$ contain enough information to recover the full span U.

In particular, we prove the following claim:

Condition 5.10 (Deterministic condition for Step 1 (c))**.** *Let the columns of two (unknown) matrices $V_1 \in \mathbb{R}^{n \times k}$ and $V_2 \in \mathbb{R}^{n \times k}$ form two basis of the same k-dimensional (unknown) subspace $\mathcal{U} \subset \mathbb{R}^n$, and let U denote an arbitrary orthonormal basis of $\mathcal{U}$. Given two s-dimensional subspaces S_1 and S_2, denote $S_3 = S_1^\perp \cup S_2^\perp$. Given two projections of $\mathcal{U}$ onto the two subspaces $S_1^\top$ and $S_2^\top$: $U_1 = Proj_{S_1^\perp} V_1$ and $U_2 = Proj_{S_2^\perp} V_2$. If $\sigma_{2s}([S_1, S_2]) > 0$ and $\sigma_k(Proj_{S_3} U) > 0$, there is an algorithm for finding $\mathcal{U}$ robustly.*

The main idea in the proof is that since s is not too large, the two subspaces $S_1^\perp$ and $S_2^\perp$ have a large intersection. Using this intersection we can "align" the two basis V_1 and V_2 and obtain $V_1^\dagger V_2$, and then it is easy to merge the two projections of the same matrix (instead of a subspace).

Moreover, we show that when applying this result to the projected span of covariance matrices, we have $s = k|\mathcal{H}| \le n/3$, and the two deterministic conditions $\sigma_{2s}([S_1, S_2]) > 0$ and $\sigma_k(\mathrm{Proj}_{S_3} V_1) > 0$ are indeed satisfied with high probability over the parameter perturbation. The detailed smoothed analysis (Lemma B.13 and B.14) and the stability analysis (Lemma B.11) are provided in Section B.3 in the appendix.

Step 2. Unfold the moments to get $\widetilde{X}_4$ and $\widetilde{X}_6$. We show that given the span of covariance matrices $\mathcal{U}$ obtained from Step 1, finding the unfolded moments $\widetilde{X}_4$, $\widetilde{X}_6$ is reduced to solving two systems of linear equations.

Recall that the challenge of recovering $\widetilde{X}_4$ and $\widetilde{X}_6$ is that the two linear mappings $\mathcal{F}_4$ and $\mathcal{F}_6$ defined in (3) are *not linearly invertible*. The key idea of this step is to make use of the span $\mathcal{U}$ to *reduce the number of variables*. Note that given the basis $U \in \mathbb{R}^{n_2 \times k}$ of the span of the covariance matrices, we can represent each vectorized covariance matrix as $\widetilde{\Sigma}^{(i)} = U\widetilde{\sigma}^{(i)}$. Now Let $\widetilde{Y}_4 \in \mathbb{R}^{k \times k}_{sym}$ and $\widetilde{Y}_4 \in \mathbb{R}^{k \times k \times k}_{sym}$ denote the unfolded moments in this new coordinate system:

$$\widetilde{Y}_4 := \sum_{i=1}^{k} \widetilde{\omega}_i \widetilde{\sigma}^{(i)} {}^{\otimes 2}, \quad \widetilde{Y}_6 = \sum_{i=1}^{k} \widetilde{\omega}_i \widetilde{\sigma}^{(i)} {}^{\otimes 3} .$$

Note that once we know $\widetilde{Y}_4$ and $\widetilde{Y}_6$, the unfolded moments $\widetilde{X}_4$ and $\widetilde{X}_6$ are given by $\widetilde{X}_4 = U\widetilde{Y}_4 U^\top$ and $\widetilde{X}_6 = \widetilde{Y}_6(U^\top, U^\top, U^\top)$. Therefore, after changing the variable, we need to solve the two linear equation systems given in (7) with the variables $\widetilde{Y}_4$ and $\widetilde{Y}_6$.

This change of variable significantly reduces the number of unknown variables. Note that the number of distinct entries in $\widetilde{Y}_4$ and $\widetilde{Y}_6$ are $k_2 = \binom{k+1}{2}$ and $k_3 = \binom{k+2}{3}$, respectively. Since $k_2 \le n_4$ and $k_3 \le n_6$, we can expect that the linear mapping from $\widetilde{Y}_4$ to $\widetilde{M}_4$ and the one from $\widetilde{Y}_6$ to $\widetilde{M}_6$ are linearly invertible. This argument is formalized below.

Condition 5.11 (Deterministic condition for Step 2)**.** *Rewrite the two systems of linear equations in (7) in their canonical form and let $\widetilde{H}_4 \in \mathbb{R}^{n_4 \times k_2}$ and $\widetilde{H}_6 \in \mathbb{R}^{n_6 \times k_3}$ denote the coefficient matrices. We can obtain the unfolded moments $\widetilde{X}_4$ and $\widetilde{X}_6$ if the coefficient matrices have full column rank.*

We show with smoothed analysis that the smallest singular value of the two coefficient matrices are lower bounded with high probability:

Lemma 5.12 (Correctness)**.** *With high probability over the parameter random perturbation, the k_2-th singular value of the coefficient matrix $\widetilde{H}_4$ is at least $\Omega(\rho^2 n/k)$, and the k_3-th singular value of the coefficient matrix $\widetilde{H}_6$ is at least $\Omega(\rho^3(n/k)^{1.5})$.*

To prove this lemma we rewrite the coefficient matrix as product of two matrices and bound their smallest singular values separately. One of the two matrices corresponds to a projection of
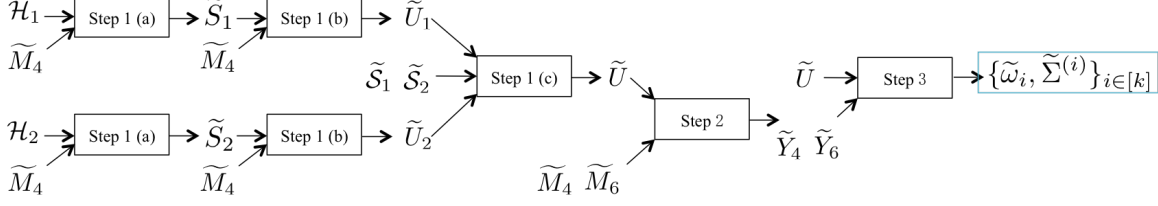
11

Figure 1: Flow of the algorithm for learning mixture of zero-mean Gaussians.

the Kronecker product $\widetilde{\Sigma} \otimes_{kr} \widetilde{\Sigma}$. In the smoothed analysis setting, this matrix is not necessarily incoherent. In order to provide a lower bound to its smallest singular value, we further apply a carefully designed projection to it, and then we use the concentration bounds for Gaussian chaoses to show that after the projection its columns are incoherent, finally we apply Gershgorin's Theorem to bound the smallest singular value [6].

When implementing this step with the empirical moments, we solve two least squares problems instead of solving the system of linear equations. Again using results in matrix perturbation theory and using the lower bound of the smallest singular values of the two coefficient matrices, we show the stability of the solution to the least squares problems:

Lemma 5.13 (Stability). *Given the empirical moments* $\widehat{M}_4 = \widetilde{M}_4 + E_4$, $\widehat{M}_6 = \widetilde{M}_6 + E_6$, *and suppose that the absolute value of entries of* E_4 *and* E_6 *are at most* δ_1. *Let* $\widehat{U}$, *the output of Step 1, be the estimation for the span of the covariance matrices, and suppose that* $\|\widehat{U} - \widetilde{U}\| \leq \delta_2$. *Let* $\widehat{Y}_4$ *and* $\widehat{Y}_6$ *be the least squares solution respectively. When* δ_1 *and* δ_2 *are inverse polynomially small, we have* $\|\widetilde{Y}_4 - \widehat{Y}_4\|_F \leq O(\sqrt{n_4}(\delta_1 + \delta_2/\sigma_{min}(\widetilde{H}_4))$ *and* $\|\widetilde{Y}_6 - \widehat{Y}_6\|_F \leq O(\sqrt{n_6}(\delta_1 + \delta_2/\sigma_{min}(\widetilde{H}_6))$.

Step 3. Tensor Decomposition.

Claim 5.14. *Given* $\widetilde{Y}_4$, $\widetilde{Y}_6$ *and* $\widetilde{U}$, *the symmetric tensor decomposition algorithm can correctly and robustly find the mixing weights* $\widetilde{\omega}_i$*'s and the vectors* $\widetilde{\sigma}_i$*'s, up to some unknown permutation over* $[k]$, *with high probability over both the randomized algorithm and the parameter perturbation.*

The algorithm and its analysis mostly follow the algorithm of symmetric tensor decomposition in Anandkumar et al. (2014), and the details are provided in Section D in the appendix.

Proof Sketch for the Main Theorem of Zero-mean Case. Theorem 3.5 follows from the previous smoothed analysis and stability analysis lemmas for each step.

First, exploiting the randomness of parameter perturbation, the smoothed analysis lemmas show that the deterministic conditions, which guarantee the correctness of each step, are satisfied with high probability. Then using concentration bounds of Gaussian variables, we show that with high probability over the random samples, the empirical moments $\widehat{M}_4$ and $\widehat{M}_6$ are entrywise δ-close to the exact moments $\widetilde{M}_4$ and $\widetilde{M}_6$. In order to achieve ϵ accuracy in the parameter estimation, we choose δ to be inverse polynomially small, and therefore the number of samples required will be polynomial in the relevant parameters. The stability lemmas show how the errors propagate only "polynomially" through the steps of the algorithm, which is visualized in Figure 1.

A more detailed illustration is provided in Section E in the appendix.

---

[6]Note that the idea of unfolding using system of linear equations also appeared in the work of Jain and Oh (2014). However, in order to show the system of linear equations in their setup is robust, i.e., the coefficient matrix has full rank, they heavily rely on the *incoherence* assumption, which we do not impose in the smoothed analysis setting.
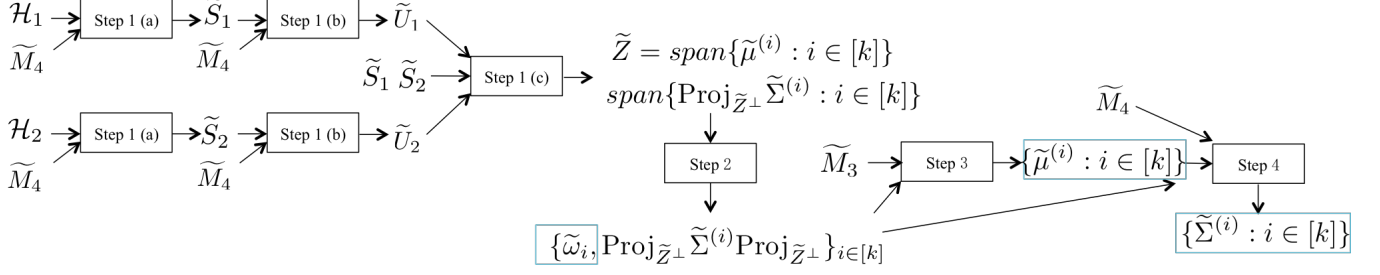
Figure 2: Flow of the algorithm for learning mixtures of general Gaussians.

6 Algorithm Outline for Learning Mixture of General Gaussians

In this section, we briefly discuss the algorithm for learning mixture of *general* Gaussians. Figure 2 shows the inputs and outputs of each step in this algorithm. Many steps share similar ideas to those of the algorithm for the zero-mean case in previous sections. We only highlight the basic ideas and defer the details to Section F in the appendix.

Step 1. Find $\widetilde{Z} = span\{\widetilde{\mu}^{(i)} : i \in [k]\}$ **and** $\widetilde{\Sigma}_o = span\{\mathbf{Proj}_{\widetilde{Z}^\perp}\widetilde{\Sigma}^{(i)}\mathbf{Proj}_{\widetilde{Z}^\perp} : i \in [k]\}.$ Similar to Step 1 in the zero-mean case, this step makes use of the structure of the 4-th order moments $\widetilde{M}_4$, and is achieved in three small steps:

(a) For a subset $\mathcal{H} \subset [n]$ of size $|\mathcal{H}| = \sqrt{n}$, find the span:

$$\mathcal{S} = \mathrm{span}\left\{\widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)}_{[:,j]} : i \in [k], j \in \mathcal{H}\right\} \subset \mathbb{R}^n. \tag{10}$$

(b) Find the span of the covariance matrices with the columns projected onto $\mathcal{S}^\perp$, namely,

$$\mathcal{U}_S = \mathrm{span}\left\{\mathrm{vec}(\mathrm{Proj}_{S^\perp}\widetilde{\Sigma}^{(i)}) : i \in [k]\right\} \subset \mathbb{R}^{n^2}. \tag{11}$$

(c) For disjoint subsets $\mathcal{H}_1$ and $\mathcal{H}_2$, repeat Step 1 (a) and Step 1 (b) to obtain $\mathcal{U}_1$ and $\mathcal{U}_2$, the span of the covariance matrices projected onto the subspaces $\mathcal{S}_1^\perp$ and $\mathcal{S}_2^\perp$. The intersection of the two subspaces $\mathcal{U}_1$ and $\mathcal{U}_2$ gives the span of the mean vectors $\widetilde{Z} = \mathrm{span}\left\{\widetilde{\mu}^{(i)}, i \in [k]\right\}$. Merge the two subspaces $\mathcal{U}_1$ and $\mathcal{U}_2$ to obtain the span of the covariance matrices projected to the subspace orthogonal to $\widetilde{Z}$, namely $\widetilde{\Sigma}_o = \mathrm{span}\left\{\mathrm{Proj}_{\widetilde{Z}^\perp}\widetilde{\Sigma}^{(i)}\mathrm{Proj}_{\widetilde{Z}^\perp} : i \in [k]\right\}.$

Step 2. Find the Covariance Matrices in the Subspace $\widetilde{Z}^\perp$ **and the Mixing Weights** $\widetilde{\omega}_i$**'s.** The key observation of this step is that when the samples are projected to the subspace orthogonal to all the mean vectors, they are equivalent to samples from a mixture of zero-mean Gaussians with covariance matrices $\widetilde{\Sigma}^{(i)}_o = \mathrm{Proj}_{\widetilde{Z}^\perp}\widetilde{\Sigma}^{(i)}\mathrm{Proj}_{\widetilde{Z}^\perp}$ and with the same mixing weights $\widetilde{\omega}_i$'s. Therefore, projecting the samples to $\widetilde{Z}^\perp$, the subspace orthogonal to the mean vectors, and use the algorithm for the zero-mean case, we can obtain $\widetilde{\Sigma}^{(i)}_o$'s, the covariance matrices projected to this subspace, as well as the mixing weights $\widetilde{\omega}_i$'s.

Step 3. Find the means With simple algebra, this step extracts the projected covariance matrices $\widetilde{\Sigma}^{(i)}_o$'s from the 3-rd order moments $\widetilde{M}_3$, the mixing weights $\widetilde{\omega}_i$ and the projected covariance matrices $\widetilde{\Sigma}^{(i)}_o$'s obtained in Step 2.

13

Step 4. Find the full covariance matrices In Step 2, we obtained $\widetilde{\Sigma}_o^{(i)}$, the covariance matrices projected to the subspace orthogonal to all the means. Note that they are equal to matrices $(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top)$ projected to the same subspace. We claim that if we can find the span of these matrices $((\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top)$'s), we can get each matrix $(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top)$, and then subtracting the known rank-one component to find the covariance matrix $\widetilde{\Sigma}^{(i)}$. This is similar to the idea of merging two projections of the same subspace in Step 1 (c) for the zero-mean case.

The idea of finding the desired span is to construct a 4-th order tensor:

$$\widetilde{M}_4' = \widetilde{M}_4 + 2\sum_{i=1}^{k} \widetilde{\omega}_i(\widetilde{\mu}^{(i)}{}^{\otimes 4}),$$

which corresponds to the 4-th order moments of a mixture of zero-mean Gaussians with covariance matrices $\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top$ and the same mixing weights $\widetilde{\omega}_i$'s. Then we can then use Step 1 of the algorithm for the zero-mean case to obtain the span of the new covariance matrices, i.e. $span\{\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top : i \in [k]\}$.

7 Conclusion

In this paper we give the first efficient algorithm for learning mixture of general Gaussians in the smoothed analysis setting. In the algorithm we developed new ways of extracting information from lower-order moment structure. This suggests that although the method of moments often involves solving systems of polynomial equations that are intractable in general, for natural models there is still hope of utilizing their special structure to obtain algebraic solution.

Smoothed analysis is a very useful way of avoiding degenerate examples in analyzing algorithms. In the analysis, we proved several new results for bounding the smallest singular values of *structured* random matrices. We believe the lemmas and techniques can be useful in more general settings.

Our algorithm uses only up to 6-th order moments. We conjecture that using higher order moments can reduce the number of dimension required to $n \geq \Omega(k^{1+\epsilon})$, or maybe even $n \geq \Omega(k^\epsilon)$.

Acknowledgements

References

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15: 2773–2832, 2014. URL http://jmlr.org/papers/v15/anandkumar14b.html.

Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. *arXiv preprint arXiv:1311.2891*, 2013.

Mikhail Belkin and Kaushik Sinha. Learning gaussian mixtures with arbitrary separation. *arXiv preprint arXiv:0907.1054*, 2009.

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

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th ACM symposium on Theory of computing*, 2014.

S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008.

Siu-On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 604–613, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2710-7. doi: 10.1145/2591796.2591848. URL http://doi.acm.org/10.1145/2591796.2591848.

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

Sanjoy Dasgupta and Leonard J Schulman. A two-round variant of em for gaussian mixtures. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 152–159. Morgan Kaufmann Publishers Inc., 2000.

Victor H de la Peña and Stephen J Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, pages 806–816, 1995.

Jon Feldman, Rocco A Servedio, and Ryan O'Donnell. Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *Learning Theory*, pages 20–34. Springer, 2006.

Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Proceedings of The 27th Conference on Learning Theory*, pages 703–725, 2014a.

Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, 2014b.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

Prateek Jain and Sewoong Oh. Learning mixtures of discrete product distributions using spectral decompositions. In *Proceedings of The 27th Conference on Learning Theory*, pages 824–856, 2014.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng. Learning and smoothed analysis. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, pages 395–404. IEEE, 2009.

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

Rafał Latała et al. Estimates of moments and tails of gaussian chaoses. *The Annals of Probability*, 34(6):2315–2331, 2006.

Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

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

H Permuter, J Francos, and H Jermyn. Gaussian mixture models of texture and colour for image database retrieval. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 3, pages III–569. IEEE, 2003.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1): 72–83, 1995.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

Gilbert W Stewart and Ji-guang Sun. *Matrix perturbation theory*. Academic press, 1990.

GW Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM review*, 19(4):634–662, 1977.

Terence Tao and Van Vu. On random±1 matrices: singularity and determinant. *Random Structures & Algorithms*, 28(1):1–23, 2006.

D Michael Titterington, Adrian FM Smith, Udi E Makov, et al. *Statistical analysis of finite mixture distributions*, volume 7. Wiley New York, 1985.

Gregory John Valiant. *Algorithmic approaches to statistical questions*. PhD thesis, University of California, Berkeley, 2012.

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

Van Vu and Ke Wang. Random weighted projections, random quadratic forms and random eigenvectors. *arXiv preprint arXiv:1306.3099*, 2013.

Daniel Zoran and Yair Weiss. Natural images, gaussian mixtures and dead leaves. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1736–1744. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4758-natural-images-gaussian-mixtures-and-dead-leaves.pdf.

A   Moment Structures

In this section we characterize the structure of the 3-rd, 4-th and 6-th moments of Gaussians mixtures.

As described in Section 3.2, the m-th order moments of the Gaussian mixture model are given by the following m-th order symmetric tensor $M \in \mathbb{R}_{sym}^{n \times \cdots \times n}$:

$$[M_m]_{j_1,\ldots,j_m} := \mathbb{E}\left[x_{j_1}\ldots x_{j_m}\right] = \sum_{i=1}^{k} \omega_i \mathbb{E}\left[y_{j_1}^{(i)}\ldots y_{j_m}^{(i)}\right], \quad \forall j_1,\ldots,j_m \in [n],$$

where $y^{(i)}$ corresponds to the n-dimensional Gaussian distribution $\mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$.

Gaussian distribution is a highly symmetric distribution, and in the zero-mean case the higher moments are well-understood by Isserlis' Theorem:

Theorem A.1 (Isserlis). *Let* $\mathbf{y} = (y_1, \ldots, y_{2t})$ *be a multivariate Gaussian random vector with mean zero and covariance* Σ, *then*

$$\mathbb{E}[y_1 \ldots y_{2t}] = \sum \prod \Sigma_{u,v},$$
$$\mathbb{E}[y_1 \ldots y_{2t-1}] = 0,$$

where the summation is taken over all distinct ways of partitioning $y_1, \ldots, y_{2t}$ *into* t *pairs, which correspond to all the perfect matchings in a complete graph. Thus there are* $(2t-1)!!$ *terms in the sum, and each summand is a product of* t *terms.*

The non-zero mean case is a direct corollary using Isserlis' Theorem and linearity of expectation.

Corollary A.2. *Let* $\mathbf{y} = (y_1, \ldots, y_t)$ *be a multivariate Gaussian random vector with mean* μ *and covariance* Σ, *then*
$$\mathbb{E}[y_1 \ldots y_t] = \sum \prod \Sigma_{u,v} \prod \mu_w.$$
where the summation is taken over all distinct ways of partitioning $y_1, \ldots, y_t$ *into* p *pairs of* (u, v) *and* s *singletons of* (w), *where* $p \geq 0$, $s \geq 0$ *and* $2p + s = t$.

As an example, $\mathbb{E}[y_1 y_2 y_3] = \mu_1 \mu_2 \mu_3 + \mu_1 \Sigma_{2,3} + \mu_2 \Sigma_{1,3} + \mu_3 \Sigma_{1,2}$.

A.1   Proof of Lemma 3.7

We shall first prove Lemma 3.7 in Section 3.2. Recall that this lemma shows that for mixture of zero-mean Gaussians, the 4-th moments $\overline{M}_4$ and the 6-th moments $\overline{M}_6$ with distinct indices can be viewed as a linear projection of the unfolded moment X_4 and X_6 defined in (1).

Proof. (of Lemma 3.7)

By Isserlis Theorem A.1, the mapping $\sqrt{3}\mathcal{F}_4$ is characterized by: $(\forall 1 \leq j_1 < j_2 < j_3 < j_4 \leq n)$

$$[M_4]_{j_1,j_2,j_3,j_4} = \sum_{i=1}^{k} \omega_i (\Sigma_{j_1,j_2}^{(i)} \Sigma_{j_3,j_4}^{(i)} + \Sigma_{j_1,j_3}^{(i)} \Sigma_{j_2,j_4}^{(i)} + \Sigma_{j_1,j_4}^{(i)} \Sigma_{j_2,j_3}^{(i)})$$
$$= [X_4]_{(j_1,j_2),(j_3,j_4)} + [X_4]_{(j_1,j_3),(j_2,j_4)} + [X_4]_{(j_1,j_4),(j_2,j_3)}.$$

Therefore, with the normalization constant $\sqrt{3}$, the (j_1, j_2, j_3, j_4)-th mapping of $\mathcal{F}_4$ is a projection of the three elements in X_4. Similarly, we have for $\sqrt{15}\mathcal{F}_6$: $(\forall 1 \le j_1 < j_2 < \cdots < j_6 \le n)$

$$
\begin{aligned}
&[M_6]_{j_1, j_2, j_3, j_4, j_5, j_6} \\
=&[X_6]_{(j_1, j_2), (j_3, j_4), (j_5, j_6)} + [X_6]_{(j_1, j_3, (j_2, j_4), (j_5, j_6)} + [X_6]_{(j_1, j_4), (j_2, j_3), (j_5, j_6)} + [X_6]_{(j_1, j_5), (j_2, j_3), (j_4, j_6)} \\
&+ [X_6]_{(j_1, j_2), (j_5, j_3), (j_4, j_6)} + [X_6]_{(j_1, j_3), (j_2, j_5), (j_4, j_6)} + [X_6]_{(j_1, j_2), (j_4, j_5), (j_3, j_6)} + [X_6]_{(j_1, j_4), (j_2, j_5), (j_3, j_6)} \\
&+ [X_6]_{(j_1, j_5), (j_2, j_4), (j_3, j_6)} + [X_6]_{(j_1, j_3), (j_4, j_5), (j_2, j_6)} + [X_6]_{(j_1, j_4), (j_3, j_5), (j_2, j_6)} + [X_6]_{(j_1, j_5), (j_3, j_2), (j_2, j_6)} \\
&+ [X_6]_{(j_2, j_3), (j_4, j_5), (j_1, j_6)} + [X_6]_{(j_2, j_4), (j_3, j_5), (j_1, j_6)} + [X_6]_{(j_2, j_5), (j_3, j_4), (j_1, j_6)}.
\end{aligned}
$$

Thus with the normalization constant $\sqrt{15}$, the mapping $\mathcal{F}_6$ is a linear projection. $\qquad\square$

A.2 Slices of Moments

Next we shall characterize the slices of the moments of mixture of Gaussians.

For mixture of zero-mean Gaussians, a one-dimensional slice of the 4th moment tensor is a vector in the span of corresponding columns of the covariance matrices:

Claim A.3 (Claim 5.1 restated)**.** *For a mixture of zero-mean Gaussians, the one-dimensional slices of the 4-th moments M_4 are given by:*

$$
M_4(e_{j_1}, e_{j_2}, e_{j_3}, I) = \sum_{i=1}^{k} \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma_{[:, j_3]}^{(i)} + \Sigma_{j_1, j_3}^{(i)} \Sigma_{[:, j_2]}^{(i)} + \Sigma_{j_2, j_3}^{(i)} \Sigma_{[:, j_1]}^{(i)} \right), \quad \forall j_1, j_2, j_3 \in [n].
$$

Proof. By the definition of multilinear map, $M_4(e_{j_1}, e_{j_2}, e_{j_3}, I)$ is a vector whose p-th entry is equal to $M_4(e_{j_1}, e_{j_2}, e_{j_3}, e_p)$. We can compute this entry by Isserlis' Theorem:

$$
M_4(e_{j_1}, e_{j_2}, e_{j_3}, e_p) = \sum_{i=1}^{k} \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma_{[p, j_3]}^{(i)} + \Sigma_{j_1, j_3}^{(i)} \Sigma_{[p, j_2]}^{(i)} + \Sigma_{j_2, j_3}^{(i)} \Sigma_{[p, j_1]}^{(i)} \right),
$$

this directly implies the claim. $\qquad\square$

For mixture of zero-mean Gaussians, a two-dimensional slice of the 4th moment M_4 is a matrix, and it is a linear combination of the covariance matrices with some additive rank one matrices:

Claim A.4 (Claim 5.6 restated)**.** *For a mixture of zero-mean Gaussians, the two-dimensional slices of the 4-th moment M_4 are given by:*

$$
M_4(e_{j_1}, e_{j_2}, I, I) = \sum_{i=1}^{k} \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma^{(i)} + \Sigma_{[:, j_1]}^{(i)} (\Sigma_{[:, j_2]}^{(i)})^\top + \Sigma_{[:, j_2]}^{(i)} (\Sigma_{[:, j_1]}^{(i)})^\top \right), \quad \forall j_1, j_2 \in [n].
$$

Proof. Again this follows from Isserlis' theorem. By definition of multilinear map this is a matrix whose (p, q)-th entry is equal to

$$
M_4(e_{j_1}, e_{j_2}, e_p, e_q) = \sum_{i=1}^{k} \omega_i \left(\Sigma_{j_1, j_2}^{(i)} \Sigma_{[p, q]}^{(i)} + \Sigma_{j_1, p}^{(i)} \Sigma_{[q, j_2]}^{(i)} + \Sigma_{j_2, p}^{(i)} \Sigma_{[q, j_1]}^{(i)} \right),
$$

and this directly implies the claim.

$\qquad\square$

Similarly, for mixture of general Gaussians, we prove the following claims:

Claim A.5 (Claim F.1 restated). *For a mixture of general Gaussians, the (j_1, j_2, j_3)-th one-dimensional slice of M_4 is given by:*

$$M_4(e_{j_1}, e_{j_2}, e_{j_3}, I) = \sum_{i=1}^{n} \omega_i \left(\mu_{j_1}^{(i)} \mu_{j_2}^{(i)} \mu_{j_3}^{(i)} \mu^{(i)} + \sum_{\pi \in \left\{ \substack{(j_1, j_2, j_3), \\ (j_2, j_3, j_1), \\ (j_3, j_1, j_2)} \right\}} \left(\Sigma_{\pi_1, \pi_2}^{(i)} \Sigma_{[:, \pi_3]}^{(i)} + \mu_{\pi_1}^{(i)} \mu_{\pi_2}^{(i)} \Sigma_{[:, \pi_3]}^{(i)} + \Sigma_{\pi_1, \pi_2}^{(i)} \mu_{\pi_3}^{(i)} \mu^{(i)} \right) \right).$$

Proof. This is very similar to Claim 5.1 and follows from the corollary of Isserlis's theorem (Corollary A.2). There are 10 ways to partition the indices $\{j_1, j_2, j_3, j_4\}$ into pairs and singletons: $((j_1), (j_2), (j_3), (j_4))$, $((j_1, j_2), (j_3), (j_4))$, $((j_1, j_3), (j_2), (j_4))$, $((j_1, j_4), (j_2), (j_3))$, $((j_2, j_3), (j_1), (j_4))$, $((j_2, j_4), (j_1), (j_3))$, $((j_3, j_4), (j_1), (j_2))$, $((j_1, j_2), (j_3, j_4))$, $((j_1, j_3), (j_2, j_4))$, $((j_1, j_4), (j_2, j_3))$. From this enumeration, we can specify each element in the vector of the one-dimensional slice. $\square$

Claim A.6 (Claim F.4 restated). *For a mixture of general Gaussians, let the matrix $M_{3(1)} \in \mathbb{R}^{n \times n^2}$ be the matricization of M_3 along the first dimension. The j-th row of $M_{3(1)}$ is given by:*

$$[M_{3(1)}]_{[j,:]} = \sum_{i=1}^{k} \omega_i \left(\mu_j^{(i)} vec(\Sigma^{(i)}) + \mu_j^{(i)} \mu^{(i)} \odot \mu^{(i)} + \Sigma_{[:,j]}^{(i)} \odot \mu^{(i)} + \mu^{(i)} \odot \Sigma_{[:,j]}^{(i)} \right)^\top.$$

Proof. Note that $[M_{3(1)}]_{[j,:]} = \left[vec(\mathbb{E}[x_j x x^\top]) \right] = vec(\mathbb{E}[x_j x \odot x])$. Again following the corollary of Isserlis's theorem (Corollary A.2, there are 4 ways to partition the indices $\{j_1, j_2, j_3\}$ into pairs and singletons: $((j_1), (j_2, j_3))$, $((j_1), (j_2), (j_3))$, $((j_1, j_2), (j_3))$, $((j_2), (j_1, j_3))$, and they correspond to the four terms in the summation.) $\square$

A.3 Two mixtures with same M_4 but different X_4

Since M_4 gives linear observations on the symmetric low rank matrix X_4, it is natural to wonder whether we can use matrix completion techniques to recover X_4 from M_4. Here we show this is impossible by giving a counter example: there are two mixture of Gaussians that generates the same 4th moment M_4, but has different X_4 (even the span of $\Sigma^{(i)}$'s are different).

By $((a, b), (c, d))$ we denote a 5×5 matrix A which has 2's on diagonals, and the only nonzero off-diagonal entries are $A_{a,b} = A_{b,a} = A_{c,d} = A_{d,c} = 1$. For example, $((1, 2), (4, 5))$ will be the following matrix:

$$\begin{pmatrix} 2 & 1 & & & \\ 1 & 2 & & & \\ & & 2 & & \\ & & & 2 & 1 \\ & & & 1 & 2 \end{pmatrix},$$

where all the missing entries are 0's. Now we construct two mixtures of 3 Gaussians, all with mean 0 and weight $1/3$. The covariance matrices are $((1, 2), (4, 5)), ((1, 3), (2, 5)), ((1, 4), (3, 5))$ for the first mixture and $((1, 2), (3, 5)), ((1, 3), (4, 5)), ((1, 4), (2, 5))$ for the second mixture. These are clearly different mixtures with different span of $\Sigma^{(i)}$'s: in the first mixture, $\Sigma_{1,2}^{(i)} = \Sigma_{4,5}^{(i)}$ for all matrices, but this is not true for the second mixture.

These two mixture of Gaussians have the same 4th moment M_4. This can be checked by using Isserlis' theorem to compute the moments. Intuitively, this is true because all the pairs $(1, i)$ and $(i, 5)$ appeared exactly twice in the covariance matrices for both mixtures; also, every 4-tuple $(1, i, j, 5)$ appeared exactly once in the covariance matrices for both mixtures.

B   Step 1: Span Finding

Recall that in Step 1 of the algorithm for learning mixture of zero-mean Gaussians, we find the span of the covariance matrices in three small steps. In this section, we prove the correctness and the robustness of each step with smoothed analysis.

For completeness we restate each substep and highlight the key properties we need, followed by the detailed proofs.

B.1   Step 1(a). Finding $\mathcal{S}$, the span of a subset of columns of $\widetilde{\Sigma}^{(i)}$'s.

Input: 4-th order moments M_4, set of indices $\mathcal{H}$.
Output: $span\{\Sigma_j^{(i)} : i \in [k], j \in \mathcal{H}\}$, represented by an orthonormal matrix $S \in \mathbb{R}^{n \times |\mathcal{H}|k}$.

Let Q be a matrix of dimension $n \times |\mathcal{H}|^3$ whose columns are all of $M_4(e_{i_1}, e_{i_2}, e_{i_3}, I)$, for $i_1, i_2, i_3 \in \mathcal{H}$.
Compute the SVD of Q: $Q = UDV^\top$.

Return: The first $k|\mathcal{H}|$ left singular vectors $S = [U_{[:,1]}, \ldots, U_{[:,k|\mathcal{H}|]}]$.

Algorithm 1: FindColumnSpan

In Step 1 (a), for any set $\mathcal{H}$ of size $\sqrt{n}$, we want to show that the one-dimensional slices of M_4 span the entire subspace $\mathcal{S} = \mathrm{span}\left\{\widetilde{\Sigma}_{[:,j]}^{(i)} : i \in [k], j \in \mathcal{H}\right\}$, which is the span of a subset of the columns in the covariance matrices.

Recall that in Claim 5.1 we showed:

$$\widetilde{M_4}(e_{j_1}, e_{j_2}, e_{j_3}, I) = \sum_{i=1}^{k} \widetilde{\omega}_i \left(\widetilde{\Sigma}_{j_1,j_2}^{(i)} \widetilde{\Sigma}_{[:,j_3]}^{(i)} + \widetilde{\Sigma}_{j_1,j_3}^{(i)} \widetilde{\Sigma}_{[:,j_2]}^{(i)} + \widetilde{\Sigma}_{j_2,j_3}^{(i)} \widetilde{\Sigma}_{[:,j_1]}^{(i)} \right), \quad \forall j_1, j_2, j_3 \in [n].$$

This in particular means when $j_1, j_2, j_3 \in \mathcal{H}$, the vector $\widetilde{M_4}(e_{j_1}, e_{j_2}, e_{j_3}, I)$ is in $\mathcal{S}$. We need to show that the columns of the matrix Q indeed span the entire subspace $\mathcal{S}$.

It is sufficient to show that a subset of the column span the entire subspace. Form a three-way even partition of the set $\mathcal{H}$, i.e., $|\mathcal{H}^{(1)}| = |\mathcal{H}^{(2)}| = |\mathcal{H}^{(3)}| = |\mathcal{H}|/3 = \sqrt{n}/3$, and only consider the one-dimensional slices of $\widetilde{M_4}$ corresponding to the indices $j_i \in \mathcal{H}^{(i)}$ for $i = 1, 2, 3$. In particular, we define matrix $\widetilde{Q}_S$ with these one-dimensional slices of $\widetilde{M_4}$:

$$\widetilde{Q}_S = \left[[[\widetilde{M_4}(e_{j_1}, e_{j_2}, e_{j_3}, I) : j_3 \in \mathcal{H}^{(3)}] : j_2 \in \mathcal{H}^{(2)}] : j_1 \in \mathcal{H}^{(1)} \right] \in \mathbb{R}^{n \times (|\mathcal{H}|/3)^3}. \tag{12}$$

Define matrix $\widetilde{P}_S$ with the corresponding columns of the covariance matrices, forming a basis (although not orthogonal) of the desired subspace $\mathcal{S}$:

$$\widetilde{P}_S = \left[[[\widetilde{\Sigma}_{[:,j]}^{(i)} : i \in [k]] : j \in \mathcal{H}^{(l)}] : l = 1, 2, 3 \right] = \left[\widetilde{\Sigma}_{[:,\mathcal{H}^{(1)}]}, \widetilde{\Sigma}_{[:,\mathcal{H}^{(2)}]}, \widetilde{\Sigma}_{[:,\mathcal{H}^{(3)}]} \right] \in \mathbb{R}^{n \times k|\mathcal{H}|}. \tag{13}$$

In the following two lemmas, we show that with high probability over the random perturbation, the column span of $\widetilde{Q}_S$ is exactly equal to the column span of $\widetilde{P}_S$, and robustly so.

Lemma B.1 (Lemma 5.3 restated). *Given $\widetilde{M_4}$, the exact 4-th order moment of the ρ-smooth mixture of zero-mean Gaussians, for any subset $\mathcal{H} \in [n]$ with cardinality $|\mathcal{H}| = \sqrt{n}$, let $\widetilde{Q}_S$ be the matrix defined as in (12) with the one-dimensional slices of $\widetilde{M_4}$. For any $\epsilon > 0$, and for some*
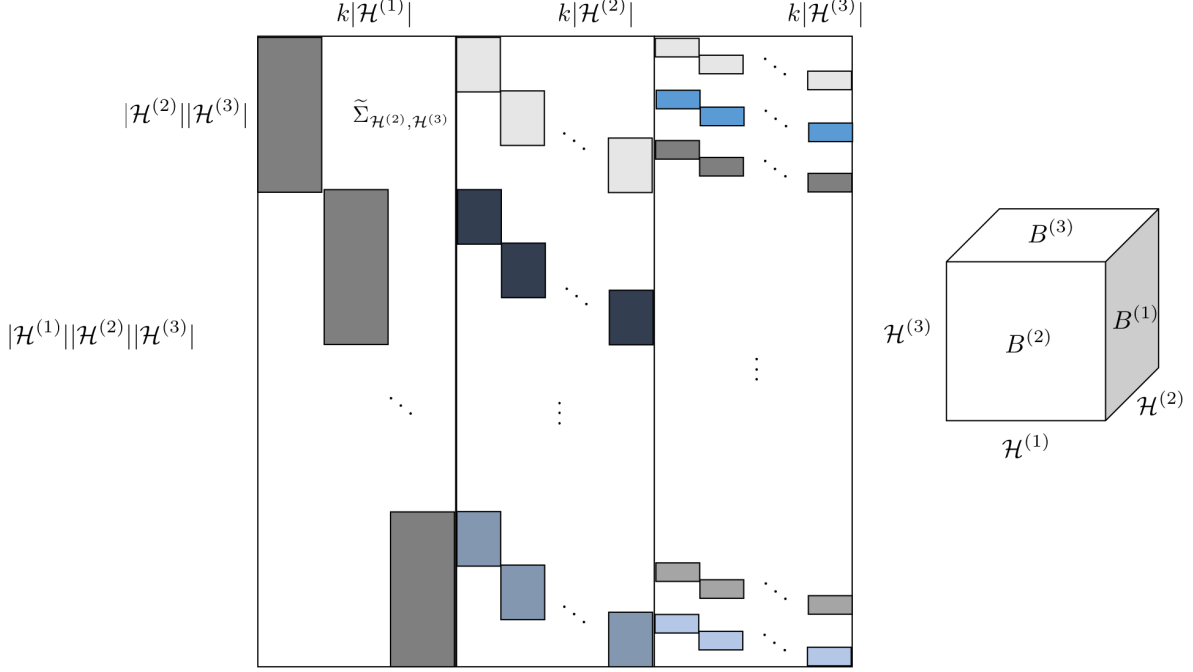
21

Figure 3: Structure of the matrix B_S

absolute constant $C_1, C_2, C_3 > 0$, with probability at least $1 - (C_1 \epsilon)^{C_2 n}$, the $k|\mathcal{H}|$-th singular value of $\widetilde{Q}_S$ is bounded below by:

$$\sigma_{k|\mathcal{H}|}(\widetilde{Q}_S) \geq C_3 \omega_o \epsilon^2 \rho^2 n. \tag{14}$$

In order to prove this lemma, we first write $\widetilde{Q}_S$ as the product of three matrices.

Claim B.2 (Structural). *Under the same assumptions of Lemma B.1, the matrix $\widetilde{Q}_S$ can be written as*

$$\widetilde{Q}_S = \widetilde{P}_S \left(D_{\widetilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|} \right) (\widetilde{B}_S)^\top, \tag{15}$$

where $\widetilde{P}_S \in \mathbb{R}^{n \times k|\mathcal{H}|}$ as defined in Equation (13 has columns equal to the columns in $\widetilde{\Sigma}^{(i)}_{[:,\mathcal{H}]}$; the diagonal matrix in the middle is the Kronecker product of two diagonal matrices and depends only on the mixing weights $\widetilde{\omega}_i$'s.

With the observation that the columns of $\widetilde{P}_S$ form a basis of the subspace $\mathcal{S}$, and each column of $\widetilde{Q}_S$ is a linear combination of the columns in $\widetilde{P}_S$, the rows of $\widetilde{B}_S \in \mathbb{R}^{(|\mathcal{H}|/3)^3 \times k|\mathcal{H}|}$ can be viewed as the coefficients for the linear combinations, and has some special structures. In particular, it consists of three blocks: $\widetilde{B}_S = \left[\widetilde{B}^{(1)}, \widetilde{B}^{(2)}, \widetilde{B}^{(3)} \right]$. The first tall matrix $\widetilde{B}^{(1)} \in \mathbb{R}^{(|\mathcal{H}|/3)^3 \times k(|\mathcal{H}|/3)}$, corresponding to the coefficient of the linear combinations on the subset of basis $\widetilde{\Sigma}_{[:,\mathcal{H}^{(1)}]}$. By the indexing order of the columns in $\widetilde{Q}_S$, the matrix $\widetilde{B}^{(1)}$ is block diagonal with identical blocks equal to $\widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}}$, defined as follows:

$$\widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}} = \left[[\widetilde{\Sigma}^{(i)}_{j_1,j_2} : j_1 \in \mathcal{H}^{(2)}, j_2 \in \mathcal{H}^{(3)}]^\top : i \in [k] \right] \in \mathbb{R}^{(|\mathcal{H}|/3)^2 \times k}.$$

With some fixed and known row permutation $\pi^{(2)}$ and $\pi^{(3)}$, the matrix $\widetilde{B}^{(2)}$ and $\widetilde{B}^{(3)}$ can be made block diagonal with identical blocks equal to $\widetilde{\Sigma}_{\mathcal{H}^{(1)},\mathcal{H}^{(3)}}$ and $\widetilde{\Sigma}_{\mathcal{H}^{(1)},\mathcal{H}^{(2)}}$, respectively. Note that the three parts $\widetilde{B}^{(1)}, \widetilde{B}^{(2)}, \widetilde{B}^{(3)}$ do not have any common entry, nor do they involve any diagonal entry of the covariance matrices, therefore the three parts are independent when the covariances are randomly perturbed in the smoothed analysis.

It is easier to understand the structure by picture, see Figure 3. The rows of the matrix should be indexed by $(j_1, j_2, j_3) \in \mathcal{H}^{(1)} \times \mathcal{H}^{(2)} \times \mathcal{H}^{(3)}$, which can also be interpreted as a cube (in the right). The block structure in the first part $\widetilde{B}^{(1)}$ correspond to a slice in $\mathcal{H}^{(2)} \times \mathcal{H}^{(3)}$ direction (for each block, the element in $\mathcal{H}^{(1)}$ is fixed, the elements in $\mathcal{H}^{(2)}$ and $\mathcal{H}^{(3)}$ take all possible values). Similarly for $\widetilde{B}^{(2)}$ and $\widetilde{B}^{(3)}$ (as shown in figure).

Proof. (of Claim B.2) The proof of this claim is using Claim 5.1, the definition of matrices and the rule of matrix multiplication. Consider the column in $\widetilde{Q}_S$ corresponding to the index (j_1, j_2, j_3) for $j_1 \in \mathcal{H}^{(1)}, j_2 \in \mathcal{H}^{(2)}, j_3 \in \mathcal{H}^{(3)}$, and the row of $\widetilde{B}_S$ together with the mixing wights specifies how this column is formed as a linear combination of $3k$ columns of $\widetilde{P}_S$. By the structure of M_4 in Claim 5.1, the (j_1, j_2, j_3)-th row of $\widetilde{B}^{(1)}$ has exactly k entries corresponding to $\widetilde{\Sigma}^{(i)}_{j_2,j_3}$ for $i \in [k]$, these entries are multiplied by $\widetilde{\omega}_i$ in the middle (diagonal) matrix. Therefore, these directly correspond to the k terms in Claim 5.1. Similarly the entries in $\widetilde{B}^{(2)}$ and $\widetilde{B}^{(3)}$ correspond to the other $2k$ terms. □

Using Claim B.2, we need to bound the smallest singular value for each of the matrices in order to bound the $k|\mathcal{H}|$-th singular value of $\widetilde{Q}_S$, this is deferred to the end of this part. The most important tool is a corollary (Lemma G.16) of the random matrix result proved in Rudelson and Vershynin (2009), which gives a lowerbound on the smallest singular value of perturbed rectangular matrices.

By Lemma B.1, we know $\widetilde{Q}_S$ has exactly rank $k|\mathcal{H}|$, and is robust in the sense that its $k|\mathcal{H}|$-th singular value is large (polynomial in the amount of perturbation ρ). By standard matrix perturbation theory, if we get $\widehat{Q}_S$ close to $\widetilde{Q}_S$ up to a high accuracy (inverse polynomial in the relevant parameters), the top $k|\mathcal{H}|$ singular vectors will span a subspace that is very close to the span of $\widetilde{Q}_S$. We formalize this in the following lemma.

Lemma B.3 (Lemma 5.4 restated). *Given the empirical estimator of the 4-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$. and suppose that the absolute value of entries of E_4 are at most δ. Let the columns of matrix $\widetilde{S} \in \mathbb{R}^{n \times k|\mathcal{H}|}$ be the left singular vector of $\widetilde{Q}_S$, and let $\widehat{S}$ be the corresponding matrix obtained with $\widehat{M}_4$. Conditioned on the high probability event $\sigma_{k|\mathcal{H}|}(\widetilde{Q}_S) > 0$, for some absolute constant C we have:*

$$\|Proj_{\widehat{S}} - Proj_{\widetilde{S}}\| \leq \frac{Cn^{1.25}}{\sigma_{k|\mathcal{H}|}(\widetilde{Q}_S)}\delta. \tag{16}$$

Proof. Note that the columns of S are the leading left singular vectors of Q_S. We apply the standard matrix perturbation bound of singular vectors. Recall that S is defined to be the first $k|\mathcal{H}|$ left singular vector of Q_S, and we have

$$\|\widehat{Q}_S - \widetilde{Q}_S\| \leq \|\widehat{Q}_S - \widetilde{Q}_S\|_F \leq \sqrt{n(|\mathcal{H}|/3)^3 \delta^2}.$$

Therefore by Wedin's Theorem (in particular the corollary Lemma G.5), we can conclude (16). □

Next, we prove Lemma B.1.

Proof of Lemma B.1   We first use Claim B.2 to write $\widetilde{Q}_S = \widetilde{P}_S \left(D_{\widetilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|} \right) (\widetilde{B}_S)^\top$, note that the matrix $(D_{\widetilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|})$ has dimension $k|\mathcal{H}| \times k|\mathcal{H}|$, therefore we just need to show with high probability each of the three factor matrix has large $k|\mathcal{H}|$-th singular value, and that implies a bound on the $k|\mathcal{H}|$-th singular value of $\widetilde{Q}_S$ by union bound. The smallest singular value of $\widetilde{P}_S$ and $\widetilde{B}_S$ are bounded below by the following two Claims.

Claim B.4. *With high probability* $\sigma_{k|\mathcal{H}|}(\widetilde{P}_S) \geq \Omega(\rho\sqrt{n})$.

Proof. This claim is easy as $\widetilde{P}_S \in \mathbb{R}^{n \times k|\mathcal{H}|}$ is a tall matrix with $n \geq 5k|\mathcal{H}|$ rows. In particular, let $\widetilde{P}_S'$ be the block of $\widetilde{P}_S$ with rows restricted to $\mathcal{H}^C = [n]\backslash\mathcal{H}$. Note that $\widetilde{P}_S'$ is a linear projection of $\mathcal{P}_S$, and by basic property of singular values in Lemma G.11, the $k|\mathcal{H}|$ singular values of $\widetilde{P}_S'$ provide lower bounds for the corresponding ones of $\widetilde{P}_S$. We only consider the restricted rows so that $\widetilde{P}_S'$ does not involve any diagonal elements of the covariance matrices, which are not randomly perturbed in our smoothed analysis framework.

Now $\widetilde{P}_S'$ is a randomly perturbed rectangular matrix, whose smallest singular value can be lower bounded using Lemma G.16, and we conclude that with probability at least $1 - (C\epsilon)^{0.25n}$,

$$\sigma_{k|\mathcal{H}|}(\widetilde{P}_S) \geq \epsilon\rho\sqrt{n}.$$

$\square$

Next, we bound the smallest singular value of $\widetilde{B}_S$.

Claim B.5. *With high probability* $\sigma_{k|\mathcal{H}|}(\widetilde{B}_S) \geq \Omega(\rho\sqrt{n})$.

Proof. We make use of the special structure of the three blocks of $\widetilde{B}_S$ to lower bound its smallest singular value.

First, we prove that the block diagonal matrix $\widetilde{B}^{(1)}$ has large singular values, even after projecting to the orthogonal subspace of the column span of $\widetilde{B}^{(2)}$ and $\widetilde{B}^{(3)}$. This idea appeared several times in our proof and is abstracted in Lemma G.12. Apply the lemma and we have:

$$\sigma_{k|\mathcal{H}|}(\widetilde{B}_S) \geq \min\left\{ \sigma_{k(2|\mathcal{H}|/3)}([\widetilde{B}^{(2)}, \widetilde{B}^{(3)}]),\ \sigma_k(\text{Proj}_{([\widetilde{B}^{(2)}, \widetilde{B}^{(3)}]_{\{j\}\times\mathcal{H}^{(2)}\times\mathcal{H}^{(3)}})^\perp}\widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}}) : j \in \mathcal{H}^{(1)}\right\} \tag{17}$$

$$\geq \min\left\{ \sigma_{k(2|\mathcal{H}|/3)}([\widetilde{B}^{(2)}, \widetilde{B}^{(3)}]),\ \sigma_k(\text{Proj}_{([\widetilde{B}^{(2)}, \widetilde{B}^{(3)}]_{\{j\}\times\mathcal{H}^{(2)}\times\mathcal{H}^{(3)}})^\perp}\text{Proj}_{\Sigma^\perp_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}}}\widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}}) : j \in \mathcal{H}^{(1)}\right\},$$

where the j-th block of $[\widetilde{B}^{(2)}, \widetilde{B}^{(3)}]$ has dimension $(|\mathcal{H}|/3)^2 \times 2k|\mathcal{H}|/3$. Since

$$(|\mathcal{H}|/3)^2 - k - 2k|\mathcal{H}|/3 = \Omega(n/9 - k - 2kn^{0.5}/3) \geq \Omega(n),$$

this means for each block, even after projection it has more than $3k$ rows. Note that by definition the three blocks $\widetilde{B}^{(1)}$, $\widetilde{B}^{(2)}$ and $\widetilde{B}^{(3)}$ are independent and do not involve any diagonal elements of the covariance matrices, so each block after the two projections is again a rectangular random matrix. We can apply Lemma G.15, for any j, for some absolute constant C_1, C_2, C_3 (not fixed throughout the discussion), with probability at least $1 - (C_1\epsilon)^{C_2 n}$ over the randomness of $\widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}}$, we have:

$$\sigma_k(\text{Proj}_{([\widetilde{B}^{(2)}, \widetilde{B}^{(3)}]_{\{j\}\times\mathcal{H}^{(2)}\times\mathcal{H}^{(3)}})^\perp}\text{Proj}_{\Sigma^\perp_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}}}\widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}}) \geq \epsilon\rho\sqrt{C_3 n}. \tag{18}$$

24

Now we can take a union bound over the blocks and conclude that with high probability, the smallest singular value of each block is large.

In order to bound $\sigma_{k(2|\mathcal{H}|/3)}([\widetilde{B}^{(2)}, \widetilde{B}^{(3)}])$, we use the same strategy. Note that $\widetilde{B}^{(2)}$ also has a block structure that corresponds to the $\mathcal{H}^{(1)} \times \mathcal{H}^{(3)}$ faces (see Figure 3). Again check the condition on dimension $(|\mathcal{H}|/3)^2 - k - k|\mathcal{H}|/3 \geq \Omega(n) > 3k$, we can apply Lemma G.12 again to show that for any j, with probability at least $1 - (C_1\epsilon)^{C_2 n}$ over the randomness of $\widetilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}$, we have:

$$\sigma_{k(2|\mathcal{H}|/3)}([\widetilde{B}^{(2)}, \widetilde{B}^{(3)}]) \geq \min\{\sigma_{k(|\mathcal{H}|/3)}(\widetilde{B}^{(3)}), \ \sigma_k(\text{Proj}_{([\widetilde{B}^{(3)}]_{\mathcal{H}^{(1)} \times \{j\} \times \mathcal{H}^{(3)}})^\perp} \text{Proj}_{\Sigma_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}^\perp} \widetilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}) : j \in \mathcal{H}^{(2)}\}. \tag{19}$$

Again by Lemma G.15, for any j, with probability at least $1 - (C_1\epsilon)^{C_2 n}$ over the randomness of $\widetilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}$, we have:

$$\sigma_k(\text{Proj}_{([\widetilde{B}^{(3)}]_{\mathcal{H}^{(1)} \times \{j\} \times \mathcal{H}^{(3)}})^\perp} \text{Proj}_{\Sigma_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}^\perp} \widetilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(3)}}) \geq \epsilon\rho\sqrt{C_3 n}. \tag{20}$$

Finally, for $\widetilde{B}^{(3)}$ it is a block diagonal structure with blocks correspond to $\mathcal{H}^{(1)} \times \mathcal{H}^{(2)}$ faces (see Figure 3). Each block is a perturbed rectangular matrix, therefore we apply Lemma G.15 to have that with high probability over the randomness of $\widetilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(2)}}$,

$$\sigma_{k(|\mathcal{H}|/3)}(\widetilde{B}^{(3)}) \geq \sigma_k(\widetilde{\Sigma}_{\mathcal{H}^{(1)}, \mathcal{H}^{(2)}}) \geq \epsilon\rho\sqrt{n}. \tag{21}$$

Now plug in the lower bounds in (18) (20) (21) into the inequalities in (17) and (19). By union bound we conclude that with high probability:

$$\sigma_{k|\mathcal{H}|}(\widetilde{B}_S) \geq \epsilon\rho\sqrt{C_3 n}.$$

$\square$

Finally, the diagonal matrix in the middle is given by the Kronecker product of $I_{|\mathcal{H}|}$ and $D_{\widetilde{\omega}}$. Recall that $D_{\widetilde{\omega}}$ is the diagonal matrix with the mixing weights $\widetilde{\omega}_i$'s on its diagonal. By property of Kronecker product and the assumption on the mixing weights, the smallest diagonal element of $D_{\widetilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|}$ is at least ω_0. Therefore $\sigma_{k|\mathcal{H}|}(D_{\widetilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|}) \geq \omega_0$.

We have shown that the smallest singular value of all the three factor matrices are large with high probability. Therefore, apply union bound, we conclude that with probability at least $1 - \exp(-\Omega(n))$,

$$\sigma_{k|\mathcal{H}|}(\widetilde{Q}_S) \geq \sigma_{k|\mathcal{H}|}(\widetilde{P}_S)\sigma_{k|\mathcal{H}|}(D_{\widetilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|})\sigma_{k|\mathcal{H}|}(\widetilde{B}_S) \geq O(\omega_o\rho^2 n).$$

B.2 Step 1 (b). Finding $\mathcal{U}_S$, the span of $\widetilde{\Sigma}^{(i)}$'s with columns projected to $\mathcal{S}^\perp$.

In Step 1 (b), given the subset of indices $\mathcal{H}$ and the subspace $\mathcal{S}$ obtained in Step 1 (a), we want to show that the projected two-dimensional slices of $\widetilde{M}_4$ span the subspace $\mathcal{U}_S$ defined in (5), which is the span of the covariance matrices with the columns projected the subspace $\mathcal{S}^\perp$:

$$\mathcal{U}_S = \text{span}\left\{\text{vec}(\text{Proj}_{\mathcal{S}^\perp}\widetilde{\Sigma}^{(i)}) : i \in [k]\right\} \subset \mathbb{R}^{n^2}.$$

Recall that in Claim 5.6, we characterized the two dimensional slices of the 4-th moments M_4 of mixture of zero-mean Gaussians as below:

$$\widetilde{M}_4(e_{j_1}, e_{j_2}, I, I) = \sum_{i=1}^k \widetilde{\omega}_i \left(\widetilde{\Sigma}_{j_1, j_2}^{(i)} \widetilde{\Sigma}^{(i)} + \widetilde{\Sigma}_{[:,j_1]}^{(i)}(\widetilde{\Sigma}_{[:,j_2]}^{(i)})^\top + \widetilde{\Sigma}_{[:,j_2]}^{(i)}(\widetilde{\Sigma}_{[:,j_1]}^{(i)})^\top\right), \quad \forall j_1, j_2 \in [n]. \tag{22}$$

25

> **Input:** 4-th order moments M_4, set of indices $\mathcal{H}$, subspace $S \subset \mathbb{R}^n$
> **Output:** $span\{\text{vec}(\text{Proj}_{S^\perp}\Sigma^{(i)}) : i \in [k]\}$, represented by an orthonormal matrix $U_S \in \mathbb{R}^{n^2 \times k}$.
>
> Let Q be a matrix whose columns are $\text{vec}(\text{Proj}_{S^\perp}M_4(e_i, e_j, I, I))$ for all $i, j \in \mathcal{H}$, $i \neq j$.
> Compute the SVD of Q: $Q = UDV^\top$.
>
> **Return:** The first k left singular vectors $U_S = [U_{[:,1]}, \ldots, U_{[:,k]}]$.

<center>**Algorithm 2:** FindProjectedSigmaSpan</center>

For notational convenience, we let $\mathcal{J}$ denote the set $\mathcal{J} = \{(j_1, j_2) : j_1 \leq j_2, \; j_1, j_2 \in \mathcal{H}\}$, and note that the cardinality is $|\mathcal{J}| = \binom{|\mathcal{H}|+1}{2} = (n + \sqrt{n})/2$. First, we define the matrix $\widetilde{Q}_{U_S} \in \mathbb{R}^{n^2 \times |\mathcal{J}|}$ whose columns are the vectorized two-dimensional slices of $\widetilde{M}_4$ with the columns projected to the subspace $\mathcal{S}^\perp$:

$$\widetilde{Q}_{U_S} = \left[\text{vec}(\text{Proj}_{S^\perp}\widetilde{M}_4(e_{j_1}, e_{j_2}, I, I)) : (j_1, j_2) \in \mathcal{J}\right]. \tag{23}$$

Similarly we define $\widetilde{Q}_{U_0} \in \mathbb{R}^{n^2 \times |\mathcal{J}|}$ with the slices without the projection:

$$\widetilde{Q}_{U_0} = \left[\text{vec}(\widetilde{M}_4(e_{j_1}, e_{j_2}, I, I)) : (j_1, j_2) \in \mathcal{J}\right].$$

Observe the structure in (22) and we see the columns of $\widetilde{Q}_{U_0}$ is "almost" in the span of covariance matrices, except for some additive rank one terms. Note that all the rank one terms lie in the subspace $\mathcal{S}$ obtained from Step 1 (a), and they vanish if we project the slice to the orthogonal subspace $\mathcal{S}^\perp$. In particular, $\text{Proj}_{\mathcal{S}^\perp}\widetilde{\Sigma}^{(i)}_{[:,j]} = 0$ for all $j \in S$. Let the columns of the matrix $\widetilde{P}_{U_S} \in \mathbb{R}^{n^2 \times k}$ be the vectorized and projected covariance matrices as below:

$$\widetilde{P}_{U_S} = \left[\text{vec}(\text{Proj}_{S^\perp}\widetilde{\Sigma}^{(i)}) : i \in [k]\right]. \tag{24}$$

In the following claim, we show that the columns of $\widetilde{Q}_{U_S}$ indeed lie in the column span of $\widetilde{P}_{U_S}$:

Claim B.6. *Given S obtained in Step 1(a), the span of $\widetilde{\Sigma}^{(i)}_{[:,j]}$ for $j \in \mathcal{H}$ and for all i, then for $j_1, j_2 \in \mathcal{H}$, we have:*

$$Proj_{S^\perp}\widetilde{M}_4(e_{j_1}, e_{j_2}, I, I) = \sum_{i=1}^{k} \widetilde{\omega}_i \widetilde{\Sigma}^{(i)}_{j_1, j_2} Proj_{S^\perp}\widetilde{\Sigma}^{(i)}, \quad \forall j_1, j_2 \in [n].$$

Similar as in Step 1(a), in the next lemma we show that the columns of $\widetilde{Q}_{U_S}$ indeed span the entire column span of $\widetilde{P}_{U_S}$. Since the dimension of the column span of $\widetilde{P}_{U_S}$ is no larger than k, it is enough to the k-th singular value of $\widetilde{Q}_{U_S}$:

Lemma B.7 (Lemma 5.8 restated). *Given $\widetilde{M}_4$, the exact 4-th order moment of the ρ-smooth mixture of Gaussians, define the matrix $\widetilde{Q}_{U_S}$ as in (23) with the two-dimensional slices of $\widetilde{M}_4$. For any $\epsilon > 0$, and for some absolute constant $C_1, C_2, C_3 > 0$, with probability at least $1 - 2(C_1\epsilon)^{C_2 n}$, the k-th singular value of $\widetilde{Q}_{U_S}$ is bounded below by:*

$$\sigma_k(\widetilde{Q}_{U_S}) \geq C_3 \omega_o(\epsilon\rho)^2 n^{1.5}.$$

<center>26</center>

Similar as before, we first examine the structure of the matrix $\widetilde{Q}_{U_S}$:

Claim B.8 (Structural). *Under the same assumption as Lemma B.7, we can write $\widetilde{Q}_{U_S}$ in the following matrix product form:*

$$\widetilde{Q}_{U_S} = \widetilde{P}_{U_S} D_{\widetilde{\omega}} \widetilde{\Sigma}_J^\top. \tag{25}$$

The columns of the matrix $\widetilde{P}_{U_S} \in \mathbb{R}^{n^2 \times k}$ are the vectorized and projected covariance matrices as defined in (24); $D_{\widetilde{\omega}}$ is the diagonal matrix with the mixing weights $\widetilde{\omega}_i$ on its diagonal; and the matrix $\widetilde{\Sigma}_J$ is defined as:

$$\widetilde{\Sigma}_J = \left[vec[\widetilde{\Sigma}^{(i)}_{(j_1,j_2)}] : (j_1, j_2) \in \mathcal{J}] : i \in [k] \right] \in \mathbb{R}^{|\mathcal{J}| \times k}.$$

Proof. This claim follows from Claim B.6, and the rule of matrix product. The coefficients $\widetilde{\omega}_i \widetilde{\Sigma}^{(i)}_{j_1,j_2}$ for the linear combinations of $\mathrm{vec}(\mathrm{Proj}_{\mathcal{S}^\perp} \widetilde{\Sigma}^{(i)})$ are given by the columns of the product $D_{\widetilde{\omega}} \widetilde{\Sigma}_J^\top$. The coefficients are then multiplied by $\widetilde{P}_{U_S}$ to select the correct columns. $\square$

To prove Lemma B.7, similar to the proof ideas of Lemma B.1, we lower bound the k-th singular value of all the three factors.

Proof of Lemma B.7 By the structural Claim B.8, we know the matrix $\widetilde{Q}_{U_S}$ can be written as a product of the three matrices as $\widetilde{Q}_{U_S} = \widetilde{P}_{U_S} D_{\widetilde{\omega}} \widetilde{\Sigma}_J^\top$.

We lower bound the k-th singular value of each of the three factors. It is easy for the last two matrices. Note that by assumption $\sigma_k(D_{\widetilde{\omega}}) \geq \omega_o$, and since $\widetilde{\Sigma}_J^\top$ is just a perturbed rectangular matrix, we can apply Lemma G.15 and with high probability we have $\sigma_k(\widetilde{\Sigma}_J) \geq \Omega(\rho\sqrt{n})$.

The first matrix $\widetilde{P}_{U_S}$ is more subtle. Let us define the projection $D_{S^\perp} = \mathrm{Proj}_{S^\perp} \otimes_{kr} I_n \in \mathbb{R}^{n^2 \times n^2}$. This is just a way of saying "apply the projection $\mathrm{Proj}_{S^\perp}$ to all columns" and then vectorize the matrix. In particular, for any matrix A we have $D_{\mathcal{S}^\perp} \mathrm{vec}(A) = \mathrm{vec}(\mathrm{Proj}_{\mathcal{S}^\perp} A)$, therefore by definition of $\widetilde{P}_{U_S}$ we can write $\widetilde{P}_{U_S} = D_{S^\perp} \widetilde{\Sigma}$.

However, we cannot apply the same trick to directly bound the smallest singular value of $D_{S^\perp}$ and $\mathrm{Proj}_{D_{S^\perp}} \widetilde{\Sigma}$ separately. The problem here is that $D_{S^\perp}$ and $\widetilde{\Sigma}$ are not independent, as the subspace S obtained in Step 1(a) also depends on the perturbation on $\widetilde{\Sigma}$, therefore $\mathrm{Proj}_{D_{S^\perp}} \widetilde{\Sigma}$ is not simply a projected perturbed matrix. Instead, we show that even conditioned on the part of randomness that is common in S and $\widetilde{\Sigma}$, $\widetilde{\Sigma}$ still has sufficient randomness due to the high dimensions, and we can still extract a tall random matrix out of it. This is elaborated in the following claim:

Claim B.9. *Under the assumptions of Lemma B.7, with high probability the matrix $\widetilde{P}_{U_S} = D_{\mathcal{S}^\perp} \widetilde{\Sigma}$ has smallest singular value at least $\Omega(\rho n)$.*

Let $\mathcal{L}$ be the set of the (j_1, j_2)-th entries of $\widetilde{\Sigma}^{(i)}$ for all i and one of j_1, j_2 is in the set $\mathcal{H}$. By Step 1(a), the subspace $\mathcal{S}' = \mathrm{span}(S, e_j : j \in \mathcal{H})$ is only dependent on the entries in $\mathcal{L}$. Here we need to include the span of e_j's for $j \in \mathcal{H}$ because the *diagonal* entries can depend on the other random perturbations. By adding the span of the vector e_j's for $j \in \mathcal{H}$ the subspace remains invariant no matter how the diagonal entries change.

Let $\mathcal{Z} = \mathrm{span}(\Sigma, S' \otimes_{kr} I_n)$, and recall that the columns of Σ are the factorization of the unperturbed covariance matrices. The subspace $\mathcal{Z}$ has dimension no larger than $|\mathcal{H}|(k+1)n + k \leq n^2/10$, and depends on the randomness of $\mathcal{L}$.

Let $\widetilde{\Sigma} = \Sigma + E$ where E is the random perturbation matrix. Now we condition on the randomness in $\mathcal{L}$. By definition the subspace $\mathcal{Z}$ is deterministic conditional on $\mathcal{L}$. However, even if we only

consider entries of $E\backslash\mathcal{L}$ there are still at least $\binom{n-k|\mathcal{H}|}{2} \geq n^2/4$ independent random variables. We shall show the randomness is enough to guarantee that the smallest singular value of $\mathrm{Proj}_{D_{\mathcal{S}^\perp}}\widetilde{\Sigma}$ is lower bounded with high probability conditioned on $\mathcal{L}$:

$$\sigma_k(\widetilde{P}_{U_S}) = \sigma_k(D_{\mathcal{S}^\perp}\widetilde{\Sigma})$$
$$\geq \sigma_k(\mathrm{Proj}_{\mathcal{Z}^\perp}\widetilde{\Sigma})$$
$$= \sigma_k(\mathrm{Proj}_{\mathcal{Z}^\perp}\Sigma + \mathrm{Proj}_{\mathcal{Z}^\perp}E)$$
$$= \sigma_k(\mathrm{Proj}_{\mathcal{Z}^\perp}E).$$

Here we used the fact that projection to a subspace cannot increase the singular values (Lemma G.11).

Conditioned on the randomness of entries in $\mathcal{L}$, $E\backslash\mathcal{L}$ still has at least $n^2/4$ random directions, while the dimension of the deterministic subspace $\mathcal{Z}$ is at most $n^2/10$. Therefore we can apply Lemma G.15 again to argue that conditionally, for every $\epsilon > 0$, with probability at least $1-(C_1\epsilon)^{C_2 n^2}$ we have:

$$\sigma_k(\widetilde{P}_{U_S}) \geq \epsilon\rho\sqrt{C_3 n^2}.$$

In summary, apply union bound and we can conclude that with probability at least $1-(C_1\epsilon)^{C_2 n}$,

$$\sigma_k(\widetilde{Q}_{U_S}) = \sigma_k(\widetilde{P}_{U_S})\sigma_k(D_{\widetilde{\omega}})\sigma_k(\widetilde{\Sigma}_J) \geq C_3 \omega_o(\epsilon\rho)^2 n^{1.5}.$$

$\square$

Next, we again use matrix perturbation bounds to prove the robustness of this step, which depends on the singular value decomposition of the matrix $\widetilde{Q}_{U_S}$.

Lemma B.10 (Lemma 5.13 restated). *Given the empirical 4-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$, and given the output $\mathrm{Proj}_{\widehat{S}^\perp}$ from Step 1 (a). Suppose that $\|\mathrm{Proj}_{\widehat{S}^\perp} - \mathrm{Proj}_{\widetilde{S}^\perp}\| \leq \delta_1$, and suppose that the absolute value of entries of E_4 are at most δ_2 for $\delta_2 \leq \|\widetilde{Q}_{U_S}\|_F/\sqrt{n^3}$. Conditioned on the high probability event $\sigma_k(\widetilde{Q}_{U_S}) > 0$, we have:*

$$\|\mathrm{Proj}_{\widehat{U}_S} - \mathrm{Proj}_{\widetilde{U}_S}\| \leq \frac{n^{2.5}\,(1+2\delta_1/\delta_2)}{\sigma_k(\widetilde{Q}_{U_S})}\delta_2. \tag{26}$$

Proof of Lemma B.10 Note that the columns of U_S are the leading left singular vectors of $\widetilde{Q}_{U_S}$. We want to apply the perturbation bound of singular vectors.

Similar to the proof of Lemma B.3, we first need to bound the spectral distance between $\widehat{Q}_{U_S}$ and $\widetilde{Q}_{U_S}$. In fact we will even bound the Frobenius norm difference:

$$\|\widehat{Q}_{U_S} - \widetilde{Q}_{U_S}\|_F = \|\widehat{D}_{S^\perp}\widehat{Q}_{U_0} - \widetilde{D}_{S^\perp}\widetilde{Q}_{U_0}\|_F$$
$$= \|\widetilde{D}_{S^\perp}(\widehat{Q}_{U_0} - \widetilde{Q}_{U_0}) + (\widehat{D}_{S^\perp} - \widetilde{D}_{S^\perp})\widetilde{Q}_{U_0} + (\widehat{D}_{S^\perp} - \widetilde{D}_{S^\perp})(\widehat{Q}_{U_0} - \widetilde{Q}_{U_0})\|_F$$
$$\leq \|\widetilde{D}_{S^\perp}\|_F\|\widehat{Q}_{U_0} - \widetilde{Q}_{U_0}\|_F + 2\|\widehat{D}_{S^\perp} - \widetilde{D}_{S^\perp}\|_F\|\widetilde{Q}_{U_0}\|_F$$
$$\leq \sqrt{n^2}\|\widetilde{D}_{S^\perp}\|_2\|\widehat{Q}_{U_0} - \widetilde{Q}_{U_0}\|_F + 2\sqrt{n}\|\mathrm{Proj}_{\widehat{S}^\perp} - \mathrm{Proj}_{\widetilde{S}^\perp}\|_F\|\widetilde{Q}_{U_0}\|_F$$
$$\leq n\sqrt{n^2|\mathcal{J}|\delta_2^2} + 2\sqrt{n}\sqrt{n^2|\mathcal{J}|}\|\mathrm{Proj}_{\widehat{S}^\perp} - \mathrm{Proj}_{\widetilde{S}^\perp}\|_F$$
$$\leq n^2\frac{|\mathcal{H}|}{\sqrt{2}}(1 + 2\|\mathrm{Proj}_{\widehat{S}^\perp} - \mathrm{Proj}_{\widetilde{S}^\perp}\|_2/\delta_2)\delta_2,$$

where we used the assumption $\|\widetilde{\Sigma}^{(i)}\| \leq 1$ to bound $\|\widetilde{Q}_{U_0}\|_F$, used the upperbound on $\|\widehat{Q}_{U_0} - \widetilde{Q}_{U_0}\|_F$ to bound the term $\|(\widehat{D}_{S^\perp} - \widetilde{D}_{S^\perp})(\widehat{Q}_{U_0} - \widetilde{Q}_{U_0})\|_F \leq \|(\widehat{D}_{S^\perp} - \widetilde{D}_{S^\perp})\|_F\delta_2\sqrt{n^2|\mathcal{J}|} \leq \|(\widehat{D}_{S^\perp} - \widetilde{D}_{S^\perp})\|_F\|\widetilde{Q}_{U_0}\|_F$, and used the fact that Frobenius norm is sub-multiplicative. Apply Wedin's Theorem (in particular the corollary Lemma G.5), we can conclude (26). $\square$
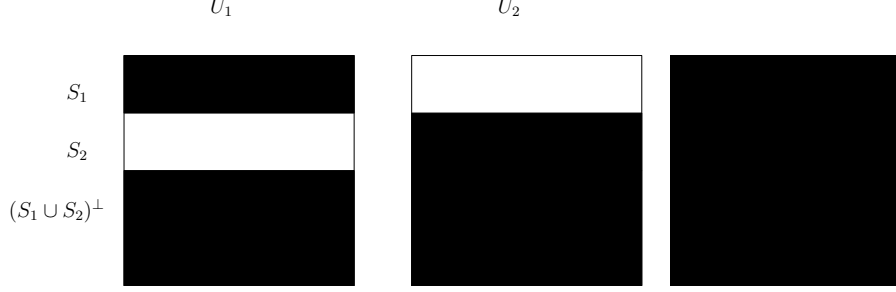
Figure 4: Step 1(c): Merging two subspaces.

B.3 Step 1 (c). Finding $\mathcal{U}$ by Merging the Two Projected Span

Input: two subspaces $S_1, S_2 \in \mathbb{R}^{n \times ks}$, two subspaces $U_1, U_2 \in \mathbb{R}^{n^2 \times k}$ (the span of covariance matrices projected to the corresponding $S_1^{\perp}, S_2^{\perp}$).
Output: $span\{\Sigma^{(i)} : i \in [k]\}$, represented by an orthonormal matrix $U \in \mathbb{R}^{n^2 \times k}$.

Let A be the first $2ks$ left singular vectors of $[S_1, S_2]$.
Let S_3 be the first $(n - 2ks)$ left singular vectors of $I - AA^{\top}$.
Let $Q = [I_{n^2}, \mathrm{Proj}_{(S_3 \otimes_{kr} I_n)} \mathrm{Proj}_{U_1}]^{\top} U_2$, compute the SVD of Q.

Return: matrix U, whose columns are the first k left singular vectors Q.
Algorithm 3: MergeProjections

Pick two disjoint sets of indices $\mathcal{H}_1, \mathcal{H}_2$, and repeat Step 1 (a) and Step 1 (b) on each of them to get $\widetilde{S}_j^{\perp}$ and $\widetilde{U}_j$ for $j = 1, 2$. In Step 1 (c), we merge the two span $\widetilde{U}_1$ and $\widetilde{U}_2$ to get $\mathcal{U}$.

If we are given two projections $\mathrm{Proj}_{S_1^{\perp}} U$ and $\mathrm{Proj}_{S_2^{\perp}} U$ of a *matrix U*, and if the union of the two subspaces $S_1^{\perp}$ and $S_2^{\perp}$ have full rank, namely $\dim(S_1 \cup S_2) = n$, then we can recover U by:

$$
U = \left[\begin{array}{c} \mathrm{Proj}_{S_1^{\perp}} \\ \mathrm{Proj}_{S_2^{\perp}} \end{array} \right]^{\dagger} \left[\begin{array}{c} \mathrm{Proj}_{S_1^{\perp}} U \\ \mathrm{Proj}_{S_2^{\perp}} U \end{array} \right].
$$

However, it is slightly different if we are given two projections of a *subspace $\mathcal{U}$*, since a subspace can be equivalently represented by different orthonormal basis up to linear transformation.

In particular, in our setting for $j = 1, 2$, we can write $\widetilde{U}_j = (\mathrm{Proj}_{S_j^{\perp}} \otimes_{kr} I_n) \widetilde{\Sigma} W_j$ for some fixed but unknown full rank matrix W_j (which makes the columns of matrix $\widetilde{\Sigma} W_j$ an orthonormal basis of $\mathcal{U}$). Recall that we define $\widetilde{\Sigma} \equiv [\mathrm{vec}(\widetilde{\Sigma}^{(i)}) : i \in [k]]$, and $D_{S_j^{\perp}} \equiv \mathrm{Proj}_{S_j^{\perp}} \otimes_{kr} I_n$ for $j = 1, 2$.

The following Lemma shows that we can still *robustly* recover the *subspace $\mathcal{U}$* if the two projections have sufficiently large overlapping. The basic idea is to use the overlapping part to align the two basis of the subspace which the two projections act on.

Lemma B.11 (Robustly merging two projections of an unknown subspace)**.** *This is the detailed statement of Condition 5.10.*

Let the columns of two fixed but unknown matrices $V_1 \in \mathbb{R}^{n \times k}$ and $V_2 \in \mathbb{R}^{n \times k}$ form two basis (not necessarily orthonormal) of the same k-dimensional fixed but unknown subspace $\mathcal{U}$ in $\mathbb{R}^n$.

For two s-dimensional known subspaces S_1 and S_2, Let the columns of A be the first $2s$ singular vectors of $[S_1, S_2]$, and let the columns of S_3 correspond to the first $(n - 2s)$ singular vectors of

29

$(I_n - Proj_A)$, therefore $S_3 \subset (S_1 \cup S_2)^\perp$. Suppose that $\sigma_k(Proj_{S_3}U) > 0$ and that $\sigma_{2s}([S_1, S_2]) > 0$. Define matrices $U_1 = Proj_{S_1^\perp}V_1$ and $U_2 = Proj_{S_2^\perp}V_2$ and we know that $U_1^\top U_1 = U_2^\top U_2 = I_k$.

We are given $\widehat{S}_1, \widehat{S}_2$ and $\widehat{U}_1, \widehat{U}_2$, and suppose that for $j = 1, 2$, we have $\|\widehat{S}_j - S_j\|_F \leq \delta_s$ and $\|\widehat{U}_j - U_j\|_F \leq \delta_u$, for $\delta_s \leq 1, \delta_u \leq 1$.

Let the columns of $\widehat{A}$ be the first $2s$ singular vectors of $[\widehat{S}_1, \widehat{S}_2]$, and let the columns of $\widehat{S}_3$ be the first $(n - 2s)$ singular vectors of $(I_n - Proj_{\widehat{A}})$. Define matrix $\widehat{U} \in \mathbb{R}^{n \times 2k}$ to be:

$$\widehat{U} = \left[\begin{array}{cc} \widehat{U}_2, & \widehat{U}_1(\widehat{S}_3^\top\widehat{U}_1)^\dagger(\widehat{S}_3^\top\widehat{U}_2) \end{array} \right] \tag{27}$$

If $\sigma_k(Proj_{S_3}U) > 0$ and $\sigma_{2s}([S_1, S_2]) > 0$, then for some absolute constant C we have:

$$\|Proj_{\widehat{U}} - Proj_U\| \leq \frac{C\sqrt{k}(\delta_u + \delta_s/\sigma_{2s}([S_1, S_2]))}{\sigma_k(Proj_{S_3}U)^2\sigma_{2s}([S_1, S_2])^3}.$$

Proof. The proof will proceed in two steps, we first show that if we are given the exact inputs, namely $\delta_s = \delta_u = 0$, then the column span of $\widehat{U}$ defined in (27) is identical to the desired subspace $\mathcal{U}$. Then we give a stability result using matrix perturbation bounds.

1. *Solving the problem using exact inputs.*

Given the exact inputs S_1, S_2, U_1, U_2, first we show that under the conditions $\sigma_{2s}([S_1, S_2]) > 0$ and $\sigma_k(Proj_{S_3}U) > 0$, then the column span of the matrix $[U_2, U_1(S_3^\top U_1)^\dagger(S_3^\top U_2)]$ is indeed identical to $\mathcal{U} = span(V_1) = span(V_2)$.

Claim B.12. *Under the same assumptions of Lemma B.11, given a matrix $V \in \mathbb{R}^{k \times k}$ such that $V = V_1^\dagger V_2$, let $Proj_{U_0}$ be the projection to the column span of $U_0 = [U_2, U_1V]$, then we have $Proj_{U_0} = Proj_U$.*

Proof. Given $V = V_1^\dagger V_2$, then $U_1V = Proj_{S_1^\perp}V_1V = Proj_{S_1^\perp}V_2$. Recall that by definition $U_2 = Proj_{S_2^\perp}V_2$, then the problem is now reduced to the simple problem of merging two projections ($U_2 = Proj_{S_2^\perp}V_2$ and $U_1V = Proj_{S_1^\perp}V_2$) of the same matrix ($V_2$). Therefore, to show that the columns of $U_0 = [U_2, U_1V]$ indeed span V_2 and thus the desired subspace U, we only need to show that $[Proj_{S_1^\perp}, Proj_{S_2^\perp}]$ has full column span. We show this by bounding the smallest singular value of it:

$$\begin{aligned}
\sigma_n([Proj_{S_2^\perp}, Proj_{S_1^\perp}]) \geq &\sigma_{2s}([Proj_{S_2^\perp}, Proj_{S_1^\perp}]\left[\begin{array}{cc} S_1 & 0 \\ 0 & S_2 \end{array} \right]) \\
= &\sigma_{2s}(\left[\begin{array}{cc} (I_n - S_2S_2^\top)S_1, & (I_n - S_1S_1^\top)S_2 \end{array} \right]) \\
= &\sigma_{2s}(\left[\begin{array}{cc} S_1, S_2 \end{array} \right]\left[\begin{array}{cc} I_s & -S_1^\top S_2 \\ -S_2^\top S_1 & I_s \end{array} \right]) \\
= &\sigma_{2s}(\left[\begin{array}{cc} S_1, S_2 \end{array} \right]\left[\begin{array}{c} S_1^\top \\ -S_2^\top \end{array} \right]\left[\begin{array}{cc} S_1, -S_2 \end{array} \right]) \\
= &\sigma_{2s}(\left[\begin{array}{cc} S_1, S_2 \end{array} \right]\left[\begin{array}{cc} S_1, -S_2 \end{array} \right]^\top\left[\begin{array}{cc} S_1, -S_2 \end{array} \right]) \\
= &\sigma_{2s}([S_1, S_2])^3 \\
> &0, \tag{28}
\end{aligned}$$

where the last inequality is by the assumption that $\sigma_{2s}([S_1, S_2]) > 0$. $\qquad \square$

Next, we show that in the exact case, the matrix $V = V_1^\dagger V_2$ can be computed by $V = (S_3^\top U_1)^\dagger (S_3^\top U_2)$. The basic idea is to use the overlapping part of the two projections U_1 and U_2 to align the two basis V_1 and V_2. Recall that by its construction, $S_3 = (S_1 \cup S_2)^\perp = S_1^\perp \cap S_2^\perp$, and $\text{Proj}_{S_3} = \text{Proj}_{S_1^\perp \cap S_2^\perp}$. Then for $j = 1$ and 2, we have:

$$S_3^\top U_j = S_3^\top \text{Proj}_{S_j^\perp} V_j = S_3^\top (\text{Proj}_{S_3} \text{Proj}_{S_j^\perp} + \text{Proj}_{S_3} \text{Proj}_{S_j^\perp}) V_j = S_3^\top (0 + \text{Proj}_{S_3}) V_j = S_3^\top V_j.$$

Moreover, since $U_j = \text{Proj}_{S_j^\perp} V_j$ is an orthonormal matrix, we have that all singular values of V_j are equal or greater than 1. Also note that U is an orthonormal matrix, so we have that $\sigma_k(\text{Proj}_{S_3} V_j) \geq \sigma_k(\text{Proj}_{S_3} U) > 0$. In other words, $S_3^\top V_j$ has full column rank k. Therefore,

$$\begin{aligned}
V &= (S_3^\top U_1)^\dagger (S_3^\top U_2) \\
&= (S_3^\top V_1)^\dagger (S_3^\top V_2) \\
&= (V_1^\top S_3 S_3^\top V_1)^{-1} V_1^\top S_3 (S_3^\top V_2) \\
&= (V_1^\top S_3 S_3^\top V_1)^{-1} V_1^\top S_3 S_3^\top V_1 V_1^\dagger V_2 \\
&= V_1^\dagger V_2
\end{aligned}$$

where the third equality is the Moore-Penrose definition, the fourth equality is because V_1 and V_2 are basis of the same subspace, there exists some full rank matrix $X \in \mathbb{R}^{k \times k}$ such that $V_2 = V_1 X$, so we have $V_1 V_1^\dagger V_2 = V_1 V_1^\dagger V_1 X = V_1 X = V_2$.

2. Stability result.

Given $\widehat{S}_1, \widehat{S}_2$ and $\widehat{U}_1, \widehat{U}_2$ which are close to the exact S_1, S_2, U_1 and U_2, we then need to bound the distance $\|\text{Proj}_{\widehat{U}} - \text{Proj}_U\|$. This follows the standard perturbation analysis. In order to apply Lemma G.5 we need to bound the distance between $\|\widehat{U} - U_0\|_F$, and lower bound the smallest singular value of U_0, namely $\sigma_k(U_0)$. Recall that we define U_0 same as in (27) for the exact case with $\delta_s = \delta_u = 0$.

First, we bound $\|\widehat{U} - U_0\|_F$. Note that we can write $U_0^\top$ as $U_0^\top = U_2 B$, where $B = [I, \quad U_1 (S_3^\top U_1)^\dagger S_3]^\top$. Recall that $S_3 = (S_1 \cup S_2)^\perp$, apply Lemma G.5 and we have:

$$\|\widehat{S}_3 - S_3\| \leq \|\text{Proj}_{\widehat{S}_1 \cup \widehat{S}_2} - \text{Proj}_{S_1 \cup S_2}\| \leq \sqrt{2} \frac{\|[\widehat{S}_1, \widehat{S}_2] - [S_1, S_2]\|_F}{\sigma_{2s}([S_1, S_2])} \leq \frac{2\sqrt{2}\delta_s}{\sigma_{2s}([S_1, S_2])}.$$

Next, note that $\|\widehat{S}_3 - S_3\| < 1$ and $\|\widehat{U}_1 - U_1\| \leq \delta_u < 1$, apply Lemma G.6 we have:

$$\|\widehat{S}_3^\top \widehat{U}_1 - S_3^\top U_1\| \leq 2(\|\widehat{S}_3 - S_3\| + \|\widehat{U}_1 - U_1\|).$$

Next, note that $\sigma_k(S_3^\top U_1) = \sigma_k(\text{Proj}_{S_3} V_1) > 0$ by assumption. Apply Lemma G.8, we have:

$$\|(\widehat{S}_3^\top \widehat{U}_1)^\dagger - (S_3^\top U_1)^\dagger\| \leq \frac{2\sqrt{2}\|\widehat{S}_3^\top \widehat{U}_1 - S_3^\top U_1\|}{\sigma_k(\text{Proj}_{S_3} V_1)^2}.$$

Next, apply Lemma G.6 again we can bound the perturbation of matrix product:

$$\begin{aligned}
\|\widehat{U} - U_0\| &= \|\widehat{U}_2 \widehat{B} - U_2 B\| \\
&\leq 2(\|\widehat{U}_2 - U_2\| + \|\widehat{B} - B\|) \\
&= 2(\|\widehat{U}_2 - U_2\| + \|\widehat{U}_1 (\widehat{S}_3^\top \widehat{U}_1)^\dagger \widehat{S}_3 - U_1 (S_3^\top U_1)^\dagger S_3\|) \\
&\leq 2(\|\widehat{U}_2 - U_2\| + 4(\|\widehat{U}_1 - U_1\| + \|(\widehat{S}_3^\top \widehat{U}_1)^\dagger - (S_3^\top U_1)^\dagger\| + \|\widehat{S}_3 - S_3\|)). \\
&\leq \frac{C(\delta_u + \delta_s / \sigma_{2s}([S_1, S_2]))}{\sigma_k(\text{Proj}_{S_3} V_1)^2},
\end{aligned}$$

where C is some absolute constant, and the last inequality summarizes the previous three inequalities, and used the fact that $\sigma_k(\text{Proj}_{S_3} V_1) < 1$. Note that $\|\widehat{U} - U_0\|_F \leq \sqrt{k}\|\widehat{U} - U_0\|$.

We are left to bound $\sigma_k(U_0)$. Recall that $\sigma_k(V_2) \geq \sigma_k(U_2) = 1$, and we have shown that in the exact case $U_0 = [\text{Proj}_{S_2^\perp} V_2, \quad \text{Proj}_{S_1^\perp} V_2]$. Then we can bound the smallest singular value of U_0 following the inequality in (28):

$$\sigma_k(U_0) \geq \sigma_n([\text{Proj}_{S_2^\perp}, \text{Proj}_{S_1^\perp}]) \geq \sigma_{2s}([S_1, S_2])^3.$$

Finally we can apply Lemma G.5 to bound the distance between the projections by:

$$\|\text{Proj}_{\widehat{U}} - \text{Proj}_{U_0}\| \leq \frac{\sqrt{2}\|\widehat{U} - U_0\|_F}{\sigma_k(U_0)} \leq \frac{C\sqrt{k}(\delta_u + \delta_s/\sigma_{2s}([S_1, S_2]))}{\sigma_k(\text{Proj}_{S_3} V_1)^2 \sigma_{2s}([S_1, S_2])^3}.$$

$\square$

In Step 1 (c), we are given the output $\widetilde{U}_1$ and $\widetilde{U}_2$ from Step 1 (b), as well as the output $\widetilde{S}_1^\perp$ and $\widetilde{S}_2^\perp$ from Step 1 (a). Recall that $\mathcal{U} = span\{\text{vec}(\widetilde{\Sigma}^{(i)}) : i \in [k]\}$, and for $j = 1, 2$, the matrix $\widetilde{U}_j$ given by Step 1 (b) corresponds to the subspace $\mathcal{U}$ projected to the subspace $\widetilde{B}_j = \widetilde{S}_j^\perp \otimes_{kr} I_n$.

Let matrix $\widetilde{S}_3 = \widetilde{S}_1^\perp \cap \widetilde{S}_2^\perp = (\widetilde{S}_1 \cup \widetilde{S}_2)^\perp$ (obtained by taking the singular vectors of $(I_n - AA^\top)$, where A corresponds to the first $2k|\mathcal{H}|$ singular vectors of $[\widetilde{S}_1, \widetilde{S}_2]$), and denote $\widetilde{B}_3 = \widetilde{S}_3 \otimes_{kr} I_n$. Define the matrix $\widetilde{Q}_U$ to be:

$$\widetilde{Q}_U = \left[\begin{array}{cc} \widetilde{U}_2, & \widetilde{U}_1(\widetilde{B}_3\widetilde{U}_1)^\dagger \widetilde{B}_3\widetilde{U}_2 \end{array} \right], \tag{29}$$

and similarly define the perturbed version $\widehat{Q}_U$ to be:

$$\widehat{Q}_U = \left[\begin{array}{cc} \widehat{U}_2, & \widehat{U}_1(\widehat{B}_3\widehat{U}_1)^\dagger \widehat{B}_3\widehat{U}_2 \end{array} \right].$$

Now we want to apply Lemma B.11 to show that $\text{Proj}_{\widetilde{Q}_U} = \text{Proj}_{\widetilde{\Sigma}}$ and bound the distance $\|\text{Proj}_{\widehat{Q}_U} - \text{Proj}_{\widetilde{\Sigma}}\|$. In order to use the lemma, we first use smoothed analysis to show (in Lemma B.13 and Lemma B.14)that the conditions required by the lemma are all satisfied with high probability over the ρ-perturbation of the covariance matrices, then conclude the robustness of Step 1 (c) in Lemma B.15.

Lemma B.13. *With high probability, for some constant C*

$$\sigma_k(Proj_{\widetilde{B}_3}\widetilde{\Sigma}) \geq C\epsilon\rho n.$$

Proof. This is in fact exactly the same as Claim B.9.

Given $\widetilde{\Sigma} = \Sigma + E$, by the definition of $\widetilde{S}_3$ and $\widetilde{B}_3$ we know that $\widetilde{B}_3$ only depends on the randomness of $P_J E$ for $i = 1, 2$, where

$$\mathcal{J} = \{(j_1, j_2) : j_1 \in \mathcal{H}_1 \cup \mathcal{H}_2, \text{ or } j_1 \in \mathcal{H}_1 \cup \mathcal{H}_2\},$$

and P_J denotes the mapping that only keeps the coordinates corresponding to the set $\mathcal{J}$. Therefore, we have:

$$\sigma_k(\text{Proj}_{\widetilde{B}_3}\widetilde{\Sigma}) \geq \sigma_k(\text{Proj}_{(\widetilde{B}_3^\top \Sigma)^\perp}\text{Proj}_{\widetilde{B}_3} E).$$

Note that the rank of $\widetilde{B}_3^\perp$ is $2nk|\mathcal{H}|$) and $|\mathcal{J}| = 2n|\mathcal{H}|$, thus $n_2 - |\mathcal{J}| - 2nk|\mathcal{H}| - k = \Omega(n^2) > 2k$. So we can apply Lemma G.15 to conclude that for some absolute constants C_1, C_2, C_3, with probability at least $1 - (C_1\epsilon)^{C_2 n^2}$, $\sigma_k(\widetilde{B}_3^\top \widetilde{\Sigma}) \geq \epsilon\rho\sqrt{C_3 n^2}$. $\square$

Lemma B.14. *With high probability, for some constant C,*

$$\sigma_{2k|\mathcal{H}|}([\widetilde{S}_1, \widetilde{S}_2]) \geq C\omega_o(\epsilon\rho)^2 n^{-0.25}.$$

Proof. For $i = 1, 2$, recall that $\widetilde{S}_i$ is the singular vectors of $\widetilde{Q}_{S_i}$, where $\widetilde{Q}_{S_i}$ is defined with the set $\mathcal{H}_i$ as in (12). We can write the singular value decomposition of $\widetilde{Q}_{S_i}$ as $\widetilde{Q}_{S_i} = \widetilde{S}_i \widetilde{D}_i \widetilde{V}_i^\top$ for some diagonal matrix $\widetilde{D}_i$ and orthonormal matrix $\widetilde{V}_i$, and

$$[\widetilde{S}_1, \widetilde{S}_2] = [\widetilde{Q}_{S_1}, \widetilde{Q}_{S_2}] \begin{bmatrix} \widetilde{V}_1 \widetilde{D}_1^{-1} & 0 \\ 0 & \widetilde{V}_2 \widetilde{D}_2^{-1} \end{bmatrix}.$$

Note that we can write $[\widetilde{Q}_{S_1}, \widetilde{Q}_{S_2}] = [\widetilde{P}_{S_1}, \widetilde{P}_{S_2}](\mathrm{diag}(B_{\widetilde{S}_1}, B_{\widetilde{S}_2}))^\top$, and following almost exactly with the proof of Lemma B.1, we can argue that, with probability at least $1 - (C_1\epsilon)^{C_2 n}$,

$$\sigma_{2k|\mathcal{H}|}([\widetilde{Q}_{S_1}, \widetilde{Q}_{S_2}]) \geq C\omega_o(\epsilon\rho)^2 n.$$

Moreover, by the structure of M_4 and the bounds on $\widetilde{\Sigma}^{(i)} \prec \frac{1}{2}I$, we can bound $\|\widetilde{Q}_{S_i}\| \leq 3\sqrt{n(|\mathcal{H}|/3)^3}$, and thus:

$$\sigma_{k|\mathcal{H}|}(V_i \widetilde{D}_i^{-1}) = \frac{1}{\sigma_{max}(\widetilde{Q}_{S_i})} \geq \frac{1}{3\sqrt{n(|\mathcal{H}|/3)^3}} = \Omega(n^{-1.25}).$$

Therefore, we can conclude that, for some absolute constant C, we have:

$$\sigma_{2k|\mathcal{H}|}([\widetilde{S}_1, \widetilde{S}_2]) \geq C\omega_o(\epsilon\rho)^2 n^{-0.25}.$$

$\square$

In the next lemma, we apply Lemma B.11 to show that under perturbation, with high probability the column span of $\mathrm{Proj}_{\widetilde{Q}_U} = \mathrm{Proj}_{\widetilde{\Sigma}}$ and this step is robust.

Lemma B.15. *Given the output $\widehat{S}_1, \widehat{S}_2$ and $\widehat{U}_1, \widehat{U}_2$ from Step 1 (a) and (b) based on the empirical moments $\widehat{M}_4$. Suppose that for $i = 1, 2$, $\|\widehat{S}_i - \widetilde{S}_i\|_F \leq \delta_s$, $\|\widehat{U}_i - \widetilde{U}_i\|_F \leq \delta_u$ for $\delta_s, \delta_u < 1$. Let the columns of $\widetilde{U} \in \mathbb{R}^{n^2 \times k}$ be the k leading singular vectors of $\widetilde{Q}_U$ defined in (29). Then for some absolute constants C, with high probability,*

$$\|\mathrm{Proj}_{\widehat{U}} - \mathrm{Proj}_{\widetilde{U}}\| \leq \frac{C\sqrt{k}(\delta_u + \delta_s n^{0.75}/(\omega_o \epsilon^2 \rho^2))}{\omega_o^3 \epsilon^8 \rho^8 n^{1.25}}. \tag{30}$$

Note that $\sigma_{2k|\mathcal{H}|n}([\widetilde{B}_1, \widetilde{B}_2]) = \sigma_{2k|\mathcal{H}|}([\widetilde{S}_1, \widetilde{S}_2])$, and for $i = 1, 2$, we have $\|\widehat{B}_i - \widetilde{B}_i\|_F \leq \sqrt{n}\|\widehat{S}_i - \widetilde{S}_i\|_F \leq \sqrt{n}\delta_s$. Therefore, with the above two smoothed analysis Lemmas showing polynomial bound of $\sigma_{2k|\mathcal{H}|}([\widetilde{S}_1, \widetilde{S}_2])$ and $\sigma_k(\mathrm{Proj}_{\widetilde{B}_3}(\widetilde{\Sigma}))$, the proof of Lemma B.15 follows by applying Lemma B.11.

C Step 2. Unfolding the Moments

In the second step of the algorithm, we solve two systems of linear equations to recover the unfolded moments.

Input: 4-th order moments $\overline{M}_4 \in \mathbb{R}^{n_4}$, 6-th order moments $\overline{M}_6 \in \mathbb{R}^{n_6}$, the span of (vectorized with distinct entries) covariance matrices $U \in \mathbb{R}^{n_2 \times k}$.
Output: Unfolded moments in the coordinate system of U: $Y_4 \in \mathbb{R}^{k \times k}_{sym}, Y_6 \in \mathbb{R}^{k \times k \times k}_{sym}$.

Let Y_4 be the solution to $\min_{Y_4 \in \mathbb{R}^{k \times k}_{sym}} \|\sqrt{3}\mathcal{F}_4(UY_4U^\top) - \overline{M}_4\|_F^2$.

Let Y_6 be the solution to $\min_{Y_6 \in \mathbb{R}^{k \times k \times k}_{sym}} \|\sqrt{15}\mathcal{F}_6 Y_6(U^\top, U^\top, U^\top) - \overline{M}_6\|_F^2$.

Return: Y_4, Y_6.

Algorithm 4: Estimate$Y_4 Y_6$

C.1   Unfolding the 4-th Order Moments

Recall the first system of linear equations is

$$\overline{M}_4 = \sqrt{3}\mathcal{F}_4 \circ \mathcal{X}_4^U(Y_4).$$

In the equation, $Y_4 \in \mathbb{R}^{k \times k}_{sym}$ is the unknown variable which can be viewed as a $k \times k$ symmetric matrix. Given $U \in \mathbb{R}^{n_2 \times k}$, the column span of $\widetilde{\Sigma}$ that we learned in Step 1, the first linear transformation $\mathcal{X}_4^U$ is simply $\mathcal{X}_4^U(Y_4) = UY_4U^\top$. It is supposed to transform Y_4 into the unfolded moments $X_4 \in \mathbb{R}^{n_2 \times n_2}_{sym}$, which is defined to be $\sum_{i=1}^k w_i \text{vec}(\widetilde{\Sigma}^{(i)})\text{vec}(\widetilde{\Sigma}^{(i)})^\top$. The next transformation $\sqrt{3}\mathcal{F}_4$ maps the unfolded moments X_4 to the folded moments $\overline{M}_4 \in \mathbb{R}^{n_4}$. As we showed in Lemma 3.7, the mapping $\mathcal{F}_4$ is a projection.

Since U is the column span matrix of $\widetilde{\Sigma}$, there must exist a Y_4 such that $X_4 = \widetilde{\Sigma}D_{\widetilde{\omega}}\widetilde{\Sigma}^\top = UY_4U^\top$ (recall that $D_{\widetilde{\omega}}$ is the diagonal matrix with entries $\widetilde{\omega}_i$), so the system must have at least one solution.

Rewrite the system of linear equations $\overline{M}_4/\sqrt{3} = \mathcal{F}_4 \circ \mathcal{X}_4^U(Y_4)$ in the canonical form: $\overline{M}_4\sqrt{3} = H_4\text{vec}(Y_4)$ where the variable $\text{vec}(Y_4) \in \mathbb{R}^{k_2}$, and the coefficient matrix $H_4 \in \mathbb{R}^{n_4 \times k_2}$ is a function of U and therefore also a function of the parameter Σ (recall $n_4 = \binom{n}{4}$ and $k_2 = \binom{k+1}{2}$). The system has a *unique* solution if the smallest singular value of the coefficient matrix H_4 is greater than zero.

The main theorem of this section shows that with high probability over the ρ-perturbation the system has a *unique* solution:

Theorem C.1. *With high probability over the ρ-perturbation of $\widetilde{\Sigma}$, the smallest singular value of the coefficient matrix $\widetilde{H}_4$ is lower bounded by $\sigma_{min}(\widetilde{H}_4) \geq \Omega(\rho^2 n/k)$. As a corollary, the system has a unique solution.*

In order to prove this theorem, we first need the following structural lemma:

Lemma C.2. *The coefficient matrix $\widetilde{H}_4$ is equal to $\widetilde{A}_4\widetilde{B}_4$. The first matrix $\widetilde{A}_4 \in \mathbb{R}^{n_4 \times k_2}$ has columns indexed by pair $\{(i,j) : 1 \leq i \leq j \leq k\}$, and the (i,j)-th column is equal to $C_{i,j}\mathcal{F}_4(vec(\widetilde{\Sigma}^{(i)})\odot vec(\widetilde{\Sigma}^{(j)}))$. Here $C_{i,j} = 1$ if $i = j$ and $C_{i,j} = 2$ if $i < j$. The second matrix $\widetilde{B}_4 \in \mathbb{R}^{k_2 \times k_2}$ transforms a $k \times k$ symmetric matrices Y_4 into:*

$$\widetilde{B}_4 vec(Y_4) = vec((\widetilde{\Sigma}^\dagger U)Y_4(\widetilde{\Sigma}^\dagger U)^\top).$$

Next we need to prove the bounds on the smallest singular values for $\widetilde{A}_4$ and $\widetilde{B}_4$. The first matrix $\widetilde{A}_4$ is essentially a projection of the Kronecker product $(\widetilde{\Sigma} \otimes_{kr} \widetilde{\Sigma})$. In particular, this projection satisfy the "symmetric off-diagonal" property defined below:

Definition C.3 (symmetric off-diagonal)**.** *Let the columns of matrix $P \in \mathbb{R}^{n_2^2 \times d_2}$ form an (arbitrary) basis of the subspace $\mathcal{P}$, and index the rows of P by pair $(i,j) \in [n_2] \times [n_2]$. The subspace $\mathcal{P}$ and the matrix P is called symmetric off-diagonal, if (i,i)-th row of P is 0 ("off-diagonal"), and the (i,j)-th row and (j,i)-th row are identical ("symmetric").*

34

Remark C.4. *Since symmetric off-diagonal is a property on the structure of rows of the basis P. If one basis of the subspace $\mathcal{P}$ is symmetric off-diagonal, then any basis is too. Moreover, any orthogonal basis of the subspace $\mathcal{P}$ will still be symmetric off-diagonal.*

Consider a Kronecker product of the same matrix $E \in \mathbb{R}^{n_2 \times k}$. The columns of $E \otimes_{kr} E$ are indexed by pair $(i,j) \in [k] \times [k]$. Consider applying a symmetric off-diagonal projection $P^\top$ to the Kronecker product. By the property of symmetry the projection will map two columns $E_{[:,i]} \odot E_{[:,j]}$ and $E_{[:,j]} \odot E_{[:,i]}$ to the same vector. Therefore the projected Kronecker product $P^\top(E \otimes_{kr} E)$ will not have full column rank k^2. However, we will show that the k_2 "unique" columns after the projection are linearly independent.

To formalize this, we define the matrix $(E \otimes_{kr} E)_{uniq} \in \mathbb{R}^{n_2^2 \times k_2}$ with the "unique" columns of $E \otimes_{kr} E$ labeled by pairs $\{(i,j) : 1 \le i \le j \le k\}$. In particular,

$$[(E \otimes_{kr} E)_{uniq}]_{[:,(i,j)]} = E_{[:,i]} \odot E_{[:,j]}.$$

In the following main lemma, we show even after projection to any symmetric off-diagonal space with sufficiently many dimensions, the "unique" columns of a Kronecker product of random matrices still has good condition number.

Lemma C.5. *Let $E \in \mathbb{R}^{n_2 \times k}$ be a Gaussian random matrix (each entry distributed as $\mathcal{N}(0,1)$). Let $P \in \mathbb{R}^{n_2^2 \times d_2}$ be a symmetric off-diagonal subspace of dimension $d_2 = \Omega(n_2^2)$. Then for any constant $C > 0$, when $n_2 \ge k^{2+C}$ we have with high probability $\sigma_{min}(P^\top(E \otimes_{kr} E)_{uniq}) \ge \Omega(n_2)$.*

Let us first see how Theorem C.1 follows from the two lemmas (Lemma C.2 and Lemma C.5).

Proof. (of Theorem C.1) Using the structural Lemma C.2, we know we only need to bound the smallest singular value of $\widetilde{A}_4$ and $\widetilde{B}_4$ separately. The following two claims directly imply the theorem.

Claim C.6. $\sigma_{min}(\widetilde{A}_4) \ge \Omega(\rho^2 n_2)$.

Claim C.7. $\sigma_{min}(\widetilde{B}_4) \ge 1/(4\|\widetilde{\Sigma}\|^2) \ge 1/(4nk)$.

Next we prove the two claims.

We apply Lemma C.5 to prove Claim C.6. Note that the ρ-perturbed covariances $\widetilde{\Sigma}$ is not a random Gaussian matrix, yet it is equal to the unperturbed matrix Σ plus a random Gaussian matrix $E_\Sigma = \rho E$[7]. Since we consider arbitrary Σ, the columns of $\widetilde{\Sigma}$ as well as the columns $\widetilde{A}_4$ may not be incoherent.

Instead, we project $\widetilde{A}_4$ to a subspace to strip away the terms involving the original matrix Σ. Let S be the range space corresponding to the projection $\mathcal{F}_4$. Recall that $|S| = n_4 = \Omega(n_2^2)$, and by the definition of $\mathcal{F}_4$, S is symmetric off-diagonal. Define the subspace $S' = \text{span}(S^\perp, \Sigma \otimes_{kr} I_{n_2}, I_{n_2} \otimes_{kr} \Sigma)$. Let $P = (S')^\perp$. By construction $|P| \ge |S| - 2kn_2 = \Omega(n_2^2)$. Also, since $P = (S')^\perp$ is a subspace of S, it must also be symmetric off-diagonal (see Remark C.4). After projecting $\widetilde{A}_4$ to P, we know that the (i,j)-th column ($1 \le i \le j \le k$) of $P^\top \widetilde{A}_4$ is given by:

$$P^\top[\widetilde{A}_4]_{[:,(i,j)]} = C_{i,j} P^\top(\Sigma_{[:,i]} \odot \Sigma_{[:,j]} + \rho E_{[:,i]} \odot \Sigma_{[:,j]} + \rho \Sigma_{[:,i]} \odot E_{[:,j]} + \rho^2 E_{[:,i]} \odot E_{[:,j]})$$
$$= C_{i,j} \rho^2 P^\top E_{[:,i]} \odot E_{[:,j]}.$$

---

[7]Note that the diagonal entries are then arbitrarily perturbed, but we will project on a symmetric off-diagonal subspace so changes on diagonal entries do not change the result.

Thus in $P^\top \widetilde{A}_4$ all the terms involving Σ disappears. Therefore

$$\sigma_{min}(\widetilde{A}_4) \geq \sigma_{min}(P^\top \widetilde{A}_4) = \sigma_{min}(P^\top(\widetilde{\Sigma} \otimes_{kr} \widetilde{\Sigma})_{uniq}) = \rho^2 \sigma_{min}(P^\top(E \otimes_{kr} E)_{uniq}) \geq \Omega(\rho^2 n_2),$$

where the first inequality is because the smallest singular value cannot become larger after projection, the first equality is by definition, the second equality is by the property of P, and the final step uses Lemma C.5[8].

For Claim C.7. Pick any $Y_4 \in \mathbb{R}^{k \times k}_{sym}$, we have

$$\|\widetilde{B}_4(Y_4)\| = \|\text{vec}((\widetilde{\Sigma}^\dagger U)Y_4(\widetilde{\Sigma}^\dagger U)^\top)\| = \|(\widetilde{\Sigma}^\dagger U)Y_4(\widetilde{\Sigma}^\dagger U)^\top\|_F \geq \|Y_4\|_F \sigma_{min}(\widetilde{\Sigma}^\dagger U)^2 = \|Y_4\|_F / \|\widetilde{\Sigma}\|^2,$$

where the inequality is because $\|AB\|_F \geq \sigma_{min}(A)\|B\|_F$ if $A \in \mathbb{R}^{m \times n}$ and $m \geq n$. Since $\|\text{vec}(Y_4)\|$ is within a factor of $\sqrt{2}$ to $\|Y_4\|_F$, and by the assumption $\widetilde{\Sigma}^{(i)} \prec \frac{1}{2}I$ we can bound $\|\widetilde{\Sigma}\| \leq \Omega(\sqrt{nk})$, we have the desired bound for $\sigma_{min}(\widetilde{B}_4)$. $\qquad \square$

Structure of the Coefficient Matrix   In this part we prove the structural Lemma C.2.

Proof. (of Lemma C.2) First, assume we know the true $\widetilde{\Sigma}$ matrix, then in order to get the unfolded moments X_4, we only need to solve the equation $\mathcal{F}_4(\widetilde{\Sigma}D_4\widetilde{\Sigma}^\top) = \overline{M}_4$ with the $k \times k$ symmetric variable D_4, and the solution should be equal to the diagonal matrix $D_{\widetilde{\omega}}$.

However, we only know U which is the column span of $\widetilde{\Sigma}$, so we can only use $UY_4U^\top$ and let $UY_4U^\top = \widetilde{\Sigma}D_4\widetilde{\Sigma}^\top$. Note that there is a one-to-one correspondence between Y_4 and D_4. In particular we know $D_4 = (\widetilde{\Sigma}^\dagger U)Y_4(\widetilde{\Sigma}^\dagger U)^\top$, this is exactly the second part $\widetilde{B}_4$.

Now the first matrix $\widetilde{A}_4$ should map $\text{vec}(D_4)$ to M_4. By construction, the (i,j)-th column $(i < j)$ of $\widetilde{A}_4$ is equal to $\mathcal{F}_4(\widetilde{\Sigma}^{(i)} \odot \widetilde{\Sigma}^{(j)} + \widetilde{\Sigma}^{(j)} \odot \widetilde{\Sigma}^{(i)}) = 2\mathcal{F}_4(\widetilde{\Sigma}^{(i)} \odot \widetilde{\Sigma}^{(j)})$, since $\mathcal{F}_4$ is symmetric off-diagonal we know $\mathcal{F}_4(v_1 \odot v_2) = \mathcal{F}_4(v_2 \odot v_1)$ for any two vectors v_1, v_2. For the (i,i)-th column, by construction they are equal to $\mathcal{F}_4(\widetilde{\Sigma}^{(i)} \odot \widetilde{\Sigma}^{(i)})$ as we wanted. $\qquad \square$

Main Lemma on Projection of Kronecker Product   In this part we prove Lemma C.5.

The singular values of Kronecker Product between two matrices are well-understood: they are just the products of the singular values of the two matrices. Therefore, the Kronecker product of two rank k matrices will have rank k^2. However, in our case the problem becomes more complicated because we only look at a projection of the resulting matrix. The projected Kronecker product may no longer have rank k^2 because of symmetry. Here we are able to show that even with projection to a low dimensional space, the rank of the new matrix is still as large as $\binom{k+1}{2}$.

The basic idea of the proof is to consider the inner-products between columns, and show that the columns are incoherent even after projection.

Proof. (of Lemma C.5) Consider the matrix $(E \otimes_{kr} E)^\top_{uniq}PP^\top(E \otimes_{kr} E)_{uniq}$, we shall show the matrix is *diagonally dominant* and hence its smallest singular value must be large. In order to do that we need to prove the following two claims:

Claim C.8. *For any $i, j \leq k$, $i \leq j$, with high probability $\|P^\top(E_{[:,i]} \odot E_{[:,j]})\|^2 \geq \Omega(n_2^2)$.*

Claim C.9. *For any $i, j \leq k$, $i \leq j$, with high probability*

$$\sum_{1 \leq i' \leq j' \leq k, (i,j) \neq (i',j')} |\langle P^\top(E_{[:,i]} \odot E_{[:,j]}), P^\top(E_{[:,i']} \odot E_{[:,j']})\rangle| \leq o(n_2^2).$$

---

[8]Note that although diagonal entries are not perturbed, we also have $P_{[i,i]} = 0$ so we can still apply the lemma.

With this two claims, we can apply Gershgorin's Disk Theorem G.9 to conclude that $\sigma_{min}((E \otimes_{kr} E)^{\top}_{uniq} PP^{\top}(E \otimes_{kr} E)_{uniq}) \geq \Omega(n_2^2)$. Therefore $\sigma_{min}(P^{\top}(E \otimes_{kr} E)_{uniq}) \geq \Omega(n_2)$.

Now we prove the two claims. For Claim C.8, it essentially says the projection of a random vector to a fixed subspace should have large norm. If the vector has independent entries, this is first shown in Tao and Vu (2006). Recently Vu and Wang (2013) generalized the result to K-concentrated vectors, see Lemma G.18. By Lemma G.19 we know conditioned on $\|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}$, $(E_{[:,i]} \odot E_{[:,j]})_{p,q}(p \neq q)$ is $O(\sqrt{n_2})$-concentrated. By assumption P ignores all the $(E_{[:,i]} \odot E_{[:,j]})_{p,p}$ entries. Therefore $\Pr[|\|P^{\top}(E_{[:,i]} \odot E_{[:,j]})\|^2 - d_2| \geq 2t\sqrt{d_2} + t^2] \leq Ce^{-\Omega(t^2/n_2)} + e^{-\Omega(n_2)}$. We then pick $t = \sqrt{d_2}/5 \geq \Omega(n_2)$, which implies $\Pr[\|P(E_{[:,i]} \odot E_{[:,j]})\|^2 \leq d_2/2] \leq Ce^{-\Omega(n_2)}$. This is what we need for Claim C.8.

For Claim C.9, we need to bound terms of the form $\langle P^{\top}(E_{[:,i]} \odot E_{[:,j]}), P^{\top}(E_{[:,i']} \odot E_{[:,j']})\rangle$. These are degree-4 Gaussian chaoses and are well-studied in Latała et al. (2006).

We break the terms according to how many of i', j' appears in i, j.

Case 1: $i', j' \notin \{i, j\}$. In this case we first randomly pick $E_{[:,i]}, E_{[:,j]}$, and condition on the high probability event that $\|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}$. In this case the inner-product can be rewritten as $\langle PP^{\top}(E_{[:,i]} \odot E_{[:,j]}), (E_{[:,i']} \odot E_{[:,j']})\rangle$, and we know $\|PP^{\top}(E_{[:,i]} \odot E_{[:,j]})\| \leq 4n_2$. Also, since P is symmetric off-diagonal we know in this degree-2 Gaussian chaos (only $E_{[:,i']}$ and $E_{[:,j']}$ are random now) there are no "diagonal" terms. Therefore the Decoupling Theorem G.23 shows without loss of generality we can assume $i' \neq j'$. Apply Theorem G.21 we know this term is bounded by $O(n_2^{1+\epsilon})$ with high probability for any $\epsilon > 0$.

Case 2: One of i', j' is in $\{i, j\}$. Without loss of generality assume $i' \in \{i, j\}$ (the other case is symmetric). Again we first randomly pick $E_{[:,i]}, E_{[:,j]}$ and condition on the high probability event that $\|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}$ (but this will also determine $E_{[:,i']}$). After the conditioning, only $E_{[:,j']}$ is still random, and the inner-product can be rewritten as $\langle \text{mat}(PP^{\top}(E_{[:,i]} \odot E_{[:,j]}))E_{[:,i']}, E_{[:,j']}\rangle$ where the fixed vector $\text{mat}(PP^{\top}(E_{[:,i]} \odot E_{[:,j]}))E_{[:,i']}$ has norm bounded by $\|PP^{\top}(E_{[:,i]} \odot E_{[:,j]})\|\|E_{[:,i']}\| \leq 8n_2^{3/2}$. By property of Gaussian with high probability the inner-product is bounded by $O(n_2^{3/2+\epsilon})$ for any $\epsilon > 0$.

Case 3: $i', j' \in \{i, j\}$. Since i', j' cannot be equal to i, j, there is only one possibility: i', j' are both equal to one of i, j and $i \neq j$. Without loss of generality assume $i' = j' = i \neq j$. We can swap i, j with i', j' and this actually becomes Case 2. By the same argument we know this term is bounded by $O(n_2^{3/2+\epsilon})$ for any $\epsilon > 0$.

There are $O(k^2)$ terms in Case 1, $O(k)$ terms in Case 2 and $O(1)$ terms in Case 3. Therefore by union bound we know the sum is bounded by $O(kn_2^{3/2+\epsilon} + k^2 n_2^{1+\epsilon})$ with high probability. Recall we are assuming $n_2 \geq k^{2+C}$ (which only requires $n \geq k^{1+C/2}$). Choose ϵ to be a small enough constant depending on C gives the result. □

C.2   Unfolding 6-th Order Moments

Recall the second system of linear equations is

$$\overline{M}_6/\sqrt{15} = \mathcal{F}_6 \circ \mathcal{X}_6^U(Y_6).$$

In the equation, $Y_6 \in \mathbb{R}^{k \times k \times k}_{sym}$ is the unknown variable which can be viewed as a $k \times k \times k$ symmetric tensro. The first linear transformation $\mathcal{X}_6^U$ transforms Y_6 into the unfolded moments $X_6 \in \mathbb{R}^{n_2 \times n_2 \times n_2}_{sym}$, which is supposed to be equal to $\sum_{i=1}^k \widetilde{w}_i \text{vec}(\widetilde{\Sigma}^{(i)})^{\otimes 3}$. The transformation is simply $X_6 = \mathcal{X}_6^U(Y_6) = Y_4(U^{\top}, U^{\top}, U^{\top})$ where $U \in \mathbb{R}^{n_2 \times k}$ is the column span of $\widetilde{\Sigma}$ that we learned in the previous section.

The next transformation $\mathcal{F}_6$ maps the unfolded moments X_6 to the folded moments $\overline{M}_6 \in \mathbb{R}^{n_6}$, which as we showed in Lemma 3.7 is a projection. Recall that $n_6 = \binom{n}{6}$.

Rewrite the system of linear equations $\overline{M}_6/\sqrt{15} = \mathcal{F}_6 \circ \mathcal{X}_6^U(Y_6)$ in the canonical form: $\overline{M}_6/\sqrt{15} = \widetilde{H}_6\mathrm{vec}(Y_6)$ where the coefficient matrix $\widetilde{H}_6 \in \mathbb{R}^{n_6 \times k_3}$ is a function of U and therefore is a function of $\widetilde{\Sigma}$ (recall $k_3 = \binom{k+2}{3}$).

The second system of linear equations tries to unfold the 6-th order moment $\overline{M}_6$ to get Y_6. Similar to Theorem C.1 the following theorem guarantees that with high probability over the perturbation the system has a unique solution.

Theorem C.10. *With high probability over the perturbation, the coefficient matrix $\widetilde{H}_6$ has smallest singular value $\sigma_{min}(\widetilde{H}_6) \geq \Omega(\rho^3(n/k)^{1.5})$. As a corollary, the system has a unique solution.*

The proof of this theorem is very similar to the proof of Theorem C.1. Here we list the important steps and highlight the differences.

As before the theorem relies on a structural lemma (Lemma C.11), and a main lemma about the symmetric off-diagonal projection of a Kronecker product of three identical matrices (Lemma C.13).

Lemma C.11. *The coefficient matrix $\widetilde{H}_6$ is equal to $\widetilde{A}_6\widetilde{B}_6$. The first matrix $\widetilde{A}_6 \in \mathbb{R}^{n_6 \times k_3}$ has columns indexed by triples (i_1, i_2, i_3) for $1 \leq i_1 \leq i_2 \leq i_3 \leq k$, and are given by:*

$$[\widetilde{A}_6]_{[:,(i_1,i_2,i_3)]} = C_{i_1,i_2,i_3}\mathcal{F}_6(vec(\widetilde{\Sigma}^{(i_1)}) \odot vec(\widetilde{\Sigma}^{(i_2)}) \odot vec(\widetilde{\Sigma}^{(i_3)})),$$

where C_{i_1,i_2,i_3} is a constant depending only on multiplicity of the indices (i_1, i_2, i_3). The second matrix $\widetilde{B}_6 \in \mathbb{R}^{k_3 \times k_3}$ transforms a $k \times k \times k$ symmetric tensor Y_6 into:

$$\widetilde{B}_6(Y_6) = Y_6((\widetilde{\Sigma}^\dagger U)^\top, (\widetilde{\Sigma}^\dagger U)^\top, (\widetilde{\Sigma}^\dagger U)^\top).$$

Before stating the main lemma, we update the definition of symmetric off-diagonal subspace.

Definition C.12. *Let the columns of matrix $P \in \mathbb{R}^{n_2^3 \times d_3}$ form a basis of a subspace $\mathcal{P}$. Index the rows of P by triples $(i_1, i_2, i_3) \in [n_2] \times [n_2] \times [n_2]$. The matrix P and the subspace $\mathcal{P}$ are called symmetric off-diagonal if: whenever i_1, i_2, i_3 are not distinct the corresponding row is 0 ("off-diagonal"); and for any permutation π over $\{1, 2, 3\}$, the rows corresponding to (i_1, i_2, i_3) and $(i_{\pi(1)}, i_{\pi(2)}, i_{\pi(3)})$ are identical ("symmetric").*

It is easy to verify that since the moments in $\overline{M}_6$ all have indices corresponding to distinct variables, the projection $\mathcal{F}_6$ is indeed symmetric off-diagonal. The constraints in this definition is closely related to the decoupling Theorem G.23 of Gaussian chaoses.

Similarly, we define the "unique" columns in the 3-way Kronecker product to be the matrix $(E \otimes_{kr} E \otimes_{kr} E)_{uniq} \in \mathbb{R}^{n_2^2 \times k_3}$ whose columns are labeled by triples $(i_1, i_2, i_3) : 1 \leq i_1 \leq i_2 \leq i_3 \leq k$, and $(E \otimes_{kr} E \otimes_{kr} E)_{uniq})_{[:,(i_1,i_2,i_3)]} = E_{[:,i_1]} \odot E_{[:,i_2]} \odot E_{[:,i_3]}$.

Lemma C.13. *Let $E \in \mathbb{R}^{n_2 \times k}$ be a Gaussian random matrix. Let $P \in \mathbb{R}^{n_2^3 \times d_3}$ be a symmetric off-diagonal subspace of dimension $d_3 \geq \Omega(n_2^3)$. For any constant $C > 0$, if $n_2 \geq k^{2+C}$, with high probability $\sigma_{min}(P^\top(E \otimes_{kr} E \otimes_{kr} E)_{uniq}) \geq \Omega(n_2^{3/2})$.*

The proofs of Theorem C.10 are based on the above two lemmas. The proof of Lemma C.11 is essentially the same as Lemma C.2. The proof of Lemma C.13 is very similar to that of Lemma C.5, and we highlight the only different case below:

Proof. (of Lemma C.13)

As before we try to prove that the columns of $P^\top(E \otimes_{kr} E \otimes_{kr} E)_{uniq}$ are incoherent. Recall we needed the following two claims:

Claim C.14. *For any $1 \le i_1 \le i_2 \le i_3 \le k$, with high probability $\|P^\top(E_{[:,i_1]} \odot E_{[:,i_2]} \odot E_{[:,i_3]})\|^2 \ge \Omega(n_2^3)$.*

Claim C.15. *For any $1 \le i_1 \le i_2 \le i_3 \le k$, with high probability*

$$\sum_{1 \le i_1' \le i_2' \le i_3', (i_1,i_2,i_3) \neq (i_1',i_2',i_3')} \left| \left\langle P^\top(E_{[:,i_1]} \odot E_{[:,i_2]} \odot E_{[:,i_3]}), P^\top(E_{[:,i_1']} \odot E_{[:,i_2']} \odot E_{[:,i_3']}) \right\rangle \right| \le o(n_2^3).$$

The first claim can still be proved by the projection Lemma G.18, except the vector $E_{[:,i_1]} \odot E_{[:,i_2]} \odot E_{[:,i_3]}$ is now $O(n_2)$-concentrated (the proof is an immediate generalization of Lemma G.19).

The second claim can be proved using similar ideas, however there is one new case. We again separate the terms according to the number of i_1', i_2', i_3' that do not appear in $\{i_1, i_2, i_3\}$.

Case 1: At least one of i_1', i_2', i_3' does not appear in $\{i_1, i_2, i_3\}$. Suppose there are t of i_1', i_2', i_3' that do not appear in $\{i_1, i_2, i_3\}$, similar to before we first sample $E_{i_1}, E_{i_2}, E_{i_3}$ and condition on the event that they all have norm at most $2\sqrt{n_2}$. The inner-product then becomes an order t Gaussian chaos with Frobenius norm $n_2^{6-t/2}$. By Theorem G.23 and Theorem G.21 we know with high probability all these terms are bounded by $n_2^{6-t/2+\epsilon}$ for any constant $\epsilon > 0$.

Case 2: All of i_1', i_2', i_3' appear in $\{i_1, i_2, i_3\}$. In the previous proof (of Lemma C.5), there was only one possibility and it reduces to Case 1. However for 6-th moment we have a new case: $i = i_1 = i_2 = i_1' < i_2' = i_3' = i_3 = j$ (and the symmetric case $i_1 = i_1' = i_2' < i_2 = i_3 = i_3'$). For this we will treat $T = PP^\top$ as a 6-th order tensor with Frobenius norm at most $n_2^{3/2}$ (as a matrix it has spectral norm 1, and rank at most n_2^3). The tensor is applied to the vectors $E_{[:,i]}$ and $E_{[:,j]}$ as $T(E_{[:,i]}, E_{[:,i]}, E_{[:,j]}, E_{[:,i]}, E_{[:,j]}, E_{[:,j]})$. First we sample $E_{[:,i]}$, by Lemma G.24 we know with high probability what remains will be a 3-rd order tensor $T(E_{[:,i]}, E_{[:,i]}, I, E_{[:,i]}, I, I)$ with Frobenius norm bounded by $O(n_2^{2+\epsilon})$. Notice that here it is important that Lemma G.24 can handle diagonal entries, because $E_{[:,i]}$ appears on the $1, 2, 4$-th coordinate (instead of the first three). We the apply Lemma G.24 again on $T(E_{[:,i]}, E_{[:,i]}, I, E_{[:,i]}, I, I)(E_{[:,j]}, E_{[:,j]}, E_{[:,j]})^9$, and conclude that with high probability the term is bounded by $O(n_2^{2.5+2\epsilon})$ which is still much smaller than n_2^3.

Finally we take the sum over all terms and choose ϵ to be small enough (depending on C), then when $k^{2+C} \le n_2$ the sum is a lower-order term. $\qquad\square$

C.3 Stability Bounds

For the two linear equation systems in (7), we can write them in canonical form with coefficient matrices $\widetilde{H}_4, \widetilde{H}_6$ and the unknown variable $\text{vec}(Y_4), \text{vec}(Y_6)$, corresponding to the k_2, k_3 distinct elements in symmetric Y_4, Y_6, namely:

$$\widetilde{H}_4 \text{vec}(Y_4) = \overline{M}_4/\sqrt{3}, \quad \widetilde{H}_6 \text{vec}(Y_6) = \overline{M}_6/\sqrt{15}.$$

When $\widehat{M}_4, \widehat{M}_6$, the empirical moment estimations for $\widetilde{M}_4, \widetilde{M}_6$, are used throughout the algorithm, both the coefficient matrices $\widetilde{H}_4, \widetilde{H}_6$ and the constant terms $\overline{M}_4, \overline{M}_6$ are affected by the noise from

---

[9] The notation might be confusing here: $T(E_{[:,i]}, E_{[:,i]}, I, E_{[:,i]}, I, I)$ is a 3rd order tensor, and we are applying it to $E_{[:,j]}, E_{[:,j]}, E_{[:,j]}$. The whole expression is equal to $T(E_{[:,i]}, E_{[:,i]}, E_{[:,j]}, E_{[:,i]}, E_{[:,j]}, E_{[:,j]})$.

empirical estimation. In practice, instead of solving systems of linear equations, we solve the least square problem:

$$\min_{Y_4 \in \mathbb{R}^{k \times k}_{sym}} \| \sqrt{3} \mathcal{F}_4 (U Y_4 U^\top) - \overline{\widehat{M}_4} \|^2, \quad \min_{Y_6 \in \mathbb{R}^{k \times k \times k}_{sym}} \| \sqrt{15} \mathcal{F}_6 Y_6 (U^\top, U^\top, U^\top) - \overline{\widehat{M}_6} \|^2. \tag{31}$$

and the solution to the least square problems are given by: $\operatorname{vec}(\widehat{Y}_4) = \widehat{H}_4^\dagger \overline{\widehat{M}_4}$ and $\operatorname{vec}(\widehat{Y}_6) = \widehat{H}_6^\dagger \overline{\widehat{M}_6}$.

Lemma C.16. *Given the empirical 4-th and 6-th order moments $\widehat{M}_4 = \widetilde{M}_4 + E_4$, $\widehat{M}_6 = \widetilde{M}_6 + E_6$, and suppose that the absolute value of entries in E_4 and E_6 are at most δ_1. Let $\widehat{U}$ be the output of Step 1 for the span of the covariance matrices, and suppose that $\|\widehat{U} - \widetilde{U}\| \leq \delta_2$. Suppose that $\delta_1 \leq \min\{\|\widetilde{M}_4\|_F / \sqrt{n_4}, \|\widetilde{M}_6\|_F / \sqrt{n_6}\}$, and $\delta_2 \leq \min\{1, \sigma_{k_2}(\widetilde{H}_4)/2, \sigma_{k_3}(\widetilde{H}_6)/2\}$. Then, conditioned on the high probability event that both $\sigma_{k_2}(\widetilde{H}_4), \sigma_{k_3}(\widetilde{H}_6)$ are bounded below, we have:*

$$\|\widehat{Y}_4 - \widetilde{Y}_4\|_F \leq O\left(\left(\delta_1 + \frac{\delta_2}{\sigma_{k_2}(\widetilde{H}_4)^2}\right)\sqrt{n_4}\right).$$

$$\|\widehat{Y}_6 - \widetilde{Y}_6\|_F \leq O\left(\left(\delta_1 + \frac{\delta_2}{\sigma_{k_3}(\widetilde{H}_6)^2}\right)\sqrt{n_6}\right).$$

Proof. We write the proof for $\widehat{Y}_4$, the proof for $\widehat{Y}_6$ is exactly the same except changing the subscripts.

Recall that the coefficient matrix $\widetilde{H}_4$ corresponds to the composition of two linear mappings $\mathcal{F}_4(U Y_4 U^\top)$ on the variable Y_4. Since we have showed that $\mathcal{F}_4$ is a projection determined by the Isserlis' Theorem and independent of the empirical estimation of the moments, we can bound the perturbation on the coefficient matrices by:

$$\|\widehat{H}_4 - \widetilde{H}_4\| \leq \|\widehat{U} \odot^2 - \widetilde{U} \odot^2\| \leq 2\|\widehat{U} - \widetilde{U}\|\|\widetilde{U}\| + \|\widehat{U} - \widetilde{U}\|_2^2 \leq 3\delta_2 \leq \|\widetilde{H}_4\|.$$

Similarly, we have $\|\widehat{H}_6 - \widetilde{H}_6\| \leq \|\widehat{U} \odot^3 - \widetilde{U} \odot^3\| \leq 7\delta_2 \leq \|\widetilde{H}_6\|$.

Therefore we can analyze the stability of the solution to the least square problems in (31) as follows:

$$\begin{aligned}
\|\operatorname{vec}(\widehat{Y}_4) - \operatorname{vec}(\widetilde{Y}_4)\| &= \left\|\widehat{H}_4^\dagger \overline{\widehat{M}_4} - \widetilde{H}_4^\dagger \overline{\widetilde{M}_4}\right\| \\
&\leq O(\|\widetilde{H}_4^\dagger\|\|\overline{\widehat{M}_4} - \overline{\widetilde{M}_4}\| + \|\widehat{H}_4^\dagger - \widetilde{H}_4^\dagger\|\|\overline{\widetilde{M}_4}\|) \\
&\leq O(\|\overline{\widehat{M}_4} - \overline{\widetilde{M}_4}\| + \|\widehat{H}_4^\dagger - \widetilde{H}_4^\dagger\|\sqrt{n_4}) \\
&\leq O\left(\sqrt{n_4}(\delta_1 + \|\widehat{H}_4^\dagger\|\|\widetilde{H}_4^\dagger\|\delta_2)\right) \\
&\leq O\left(\sqrt{n_4}(\delta_1 + \frac{1}{\sigma_{k_2}(\widetilde{H}_4)^2}\delta_2)\right),
\end{aligned}$$

where the first inequality is by applying Lemma G.6 and note that $\|(\overline{\widehat{M}_4} - \overline{\widetilde{M}_4})\|_F \leq \delta_1 \sqrt{n_4} \leq \|\overline{\widetilde{M}_4}\|_F$, the second inequality is because $\|\overline{\widetilde{M}_4}\|_F \leq O(\sqrt{n_4})$, the third inequality is by applying the perturbation bound of pseudo-inverse in Theorem G.7, the fourth inequality is by the assumption that δ_2 is sufficiently small compared to the smallest singular value of $\widetilde{H}_4$ thus $\sigma_{k_2}(\widehat{H}_4) = O(\sigma_{k_2}(\widetilde{H}_4))$. $\qquad\square$

Input: the span of covariance matrices $U \in \mathbb{R}^{n_2 \times k}$ (vectorized with distinct entries), the unfolded 4-th and 6-th moments $Y_4 \in \mathbb{R}^{k \times k}$ and $Y_6 \in \mathbb{R}^{k \times k \times k}$ in the coordinate system of U.

Output: Parameters $\mathcal{G} = \{(\omega_i, \Sigma^{(i)}) : i \in [k]\}$.

Compute the SVD of Y_4: $Y_4 = V_2 \Lambda_2 V_2^\top$.

Let $G = Y_6(V_2 \Lambda_2^{-1/2}, V_2 \Lambda_2^{-1/2}, V_2 \Lambda_2^{-1/2})$

Find the (unique) first k orthogonal eigenvectors v_i and the corresponding eigenvalues λ_i of G, denoted by $\{(v_i, \lambda_i) : i \in [k]\}$

For all $i \in [k]$, let $\mathrm{vec}(\Sigma^{(i)}) = \lambda_i U V_2 \Lambda_2^{1/2} v_i$, let $\omega_i = (\lambda_i)^{-2}$.

Return: $\mathcal{G} = \{(\omega_i, \Sigma^{(i)}) : i \in [k]\}$.

Algorithm 5: TensorDecomp

D   Step 3: Tensor Decomposition

Given the estimations of the unfolded moments Y_4 and Y_6 from Step 2, and given the span of covariance matrices U from Step 1, Step 3 use tensor decomposition to robustly find the parameters of the mixture of zero-mean Gaussians.

Recall that in the coordinate system with basis U, the covariance matrices (vectorized with distinct entries) are given by $\widetilde{\Sigma}^{(i)} = \widetilde{U}\widetilde{\sigma}^{(i)}$ for all i. The unfolded moments in the same coordinate system are:

$$\widetilde{Y}_4 = \sum_{i=1}^k \widetilde{\omega}_i \widetilde{\sigma}^{(i)\otimes 2}, \quad \widetilde{Y}_6 = \sum_{i=1}^k \widetilde{\omega}_i \widetilde{\sigma}^{(i)\,\otimes 3}.$$

We will apply tensor decomposition algorithm to find the $\widetilde{\sigma}^{(i)}$'s. We restate the theorem for orthogonal symmetric tensor decomposition in Anandkumar et al. Anandkumar et al. (2014) below:

Theorem D.1 (Theorem 5.1 in Anandkumar et al. (2014)). *Consider k orthonormal vector $v_1, \ldots v_k \in \mathbb{R}^n$'s and k positive weights $\lambda_1, \ldots \lambda_k$. Define the tensor $T = \sum_{i=1}^k \lambda_i v_i^{\otimes 3}$. Given $\widehat{T} = T + E$ and assume that $\|E\| \le C_1 \min\{\lambda_i\}/k$, then there is an algorithm that finds λ_i's and v_i's in polynomial running time with the following guarantee: with probability at least $1 - e^{-n}$, for some permutation π over $[k]$ and for all $i \in [k]$, we have:*

$$\|v_i - \widehat{v}_i\| \le O(\|E\|/\lambda_i), \quad |\lambda_i - \widehat{\lambda}_i| \le O(\|E\|).$$

In order to reduce our problem to the orthogonal tensor decomposition so that the tensor power method (Algorithm 1, page 21 in Anandkumar et al. (2014)) can be applied, we use the same "whitening" technique as in Anandkumar et al. (2014). We first compute the SVD of the unfolded 4-th moments $\widetilde{Y}_4 = \widetilde{V}_2 \widetilde{\Lambda}_2 \widetilde{V}_2^\top$, then use the singular vectors to transform the unfolded 6-th moments Y_6 into an orthogonal symmetric tensor $\widetilde{Y}_6(\widetilde{V}_2 \widetilde{\Lambda}_2^{-1/2}, \widetilde{V}_2 \widetilde{\Lambda}_2^{-1/2}, \widetilde{V}_2 \widetilde{\Lambda}_2^{-1/2})$.

Next we complete the stability analysis for the two-step procedure, i.e. whitening and orthogonal tensor decomposition, which was not analyzed in Anandkumar et al. (2014).

Theorem D.2. *Consider k linearly independent vectors $a_1, \ldots, a_k \in \mathbb{R}^n$, and k positive weights $\omega_1, \ldots, \omega_k$. Define $G_2 = \sum_{i=1}^k \omega_i a_i \otimes a_i \in \mathbb{R}^{n \times n}_{sym}$ and $G_3 = \sum_{i=1}^k \omega_i a_i \otimes a_i \otimes a_i \in \mathbb{R}^{n \times n \times n}_{sym}$. Let*

$\gamma_{min} = \min\{\sigma_{min}(G_2), 1\}$, $\gamma_{\max} = \sigma_{max}(G_2)$, and let $\omega_o = \min\{\omega_i\}$. Given $\widehat{G}_2, \widehat{G}_3$ and assume that:

$$\|\widehat{G}_2 - G_2\|_F \le \delta_2 \le o\left(\frac{\gamma_{min}^{2.5}}{k\|G_3\|}\right), \quad \|\widehat{G}_3 - G_3\|_F \le \delta_3 \le o\left(\frac{\gamma_{min}^{1.5}}{k}\right).$$

There exists an algorithm that finds $\widehat{a}_i$ and $\widehat{\omega}_i$ in polynomial (in variables $(n, k, 1/\sigma_{min}(G_2))$) running time with the following guarantee: with probability at least $1 - e^{-n}$, for some permutation π over $[k]$ and for all $i \in [k]$ we have:

$$\|\widehat{a}_{\pi(i)} - a_{\pi(i)}\| \le poly(\|G_3\|, 1/\sigma_{min}(G_2), 1/\omega_o)\delta_2 + poly(\|G_3\|, 1/\sigma_{min}(G_2), 1/\omega_o)\delta_3,$$
$$\|\widehat{\omega}_i - \omega_i\| \le poly(\|G_3\|, 1/\sigma_{min}(G_2))\delta_2 + poly(\|G_3\|, 1/\sigma_{min}(G_2))\delta_3.$$

Proof. (to Theorem D.2)

1. Algorithm

We first apply the whitening technique in Anandkumar et al. (2014): Let $\widehat{G}_2 = \widehat{V}_2 \widehat{\Lambda}_2 \widehat{V}_2^\top$ be the singular value decomposition of $\widehat{G}_2$, and note that the matrix $\widehat{V}_2 \widehat{\Lambda}_2^{-1/2}$ whitens G_2 in the sense that $\widehat{G}_2(\widehat{V}_2 \widehat{\Lambda}_2^{-1/2}, \widehat{V}_2 \widehat{\Lambda}_2^{-1/2}) = I_n$. Similarly we can whiten $\widehat{G}_3$ with the matrix $\widehat{V}_2 \widehat{\Lambda}_2^{-1/2}$ and obtain the following symmetric 3-rd order tensor $\widehat{G} \in \mathbb{R}_{sym}^{k \times k \times k}$:

$$\widehat{G} = \widehat{G}_3(\widehat{V}_2 \widehat{\Lambda}_2^{-1/2}, \widehat{V}_2 \widehat{\Lambda}_2^{-1/2}, \widehat{V}_2 \widehat{\Lambda}_2^{-1/2}).$$

Note th at in the exact case with G_2 and G_3, we have that:

$$G = \sum_{i=1}^k \lambda_i v_i \otimes^3,$$

where $\lambda_i = \omega_i^{-1/2}$, and the vectors $v_i = \lambda_i^{-1} V_2^\top \Lambda_2^{-1/2} a_i$ and they are orthonormal. Also note that $\lambda_{min} \ge 1$ and $\lambda_{max} \le \omega_o^{-1/2}$. We can then apply *orthogonal tensor decomposition* (Algorithm 1 in Anandkumar et al. (2014)) to $\widehat{G}$ to robustly obtain estimations of v_i's and λ_i's. After obtaining the estimation $\widehat{v}_i$ and $\widehat{\lambda}_i$'s, we can further obtain the estimation of a_i's and ω_i's as:

$$\widehat{a}_i = \widehat{V}_2 \widehat{\Lambda}_2^{1/2} \widehat{v}_i \widehat{\lambda}_i, \quad \widehat{\omega}_i = (\widehat{\lambda}_i)^{-2} \tag{32}$$

2. Stability analysis

The estimation of the vectors and weights are given in (32). In order to bound the distance $\|\widehat{a}_i - a_i\|$ and $\|\widehat{\omega}_i - \omega_i\|$, we show the stability of the estimation $\widehat{V}_2$, $\widehat{\Lambda}_2$, and $\widehat{v}_i$, $\widehat{\lambda}_i$ separately.

First, note that by assumption $\|\widehat{G}_2 - G_2\|_F \le \delta_2$, we can apply Lemma G.2 and Lemma G.3 to bound the singular values and the singular vectors of $\widehat{G}_2$ by:

$$\|\widehat{V}_2 - V_2\| \le \sqrt{2}\delta_2/\gamma_{min}, \quad \|\widehat{\Lambda}_2 - \Lambda_2\| \le \delta_2.$$

Define $X = V_2 \Lambda_2^{-1/2}$ and define $\Delta_X = \widehat{X} - X$. By the assumption that $\delta_2 \le o(\gamma_{min})$, we have $\|\widehat{V}_2 - V_2\| \le 1$ and $\|\widehat{\Lambda}_2^{-1/2} - \Lambda_2^{-1/2}\| \le \|\Lambda_2^{-1/2}\| \le \gamma_{min}^{-1/2}$. Therefore we can apply Lemma G.6 to bound $\|\Delta_X\|$:

$$\|\Delta_X\| \le O(\|\widehat{V}_2 - V_2\|\|\Lambda_2^{-1/2}\| + \|V_2\|\|\widehat{\Lambda}_2^{-1/2} - \Lambda_2^{-1/2}\|)$$
$$\le O\left(\frac{\delta_2}{\gamma_{min}^1}\gamma_{min}^{-1/2} + (\gamma_{min}^{-1/2})^2\delta_2\right)$$
$$\le O(\delta_2/\gamma_{min}^{1.5}.)$$

Moreover, since $\delta_2 \leq o(\gamma_{min})$, we also have $\|\Delta_X\| \leq \|X\| = \gamma_{min}^{-0.5}$.

Next, we bound the distance $\|\widehat{G} - G\|$. Recall that $\widehat{G} = \widehat{G}_3(\widehat{X}, \widehat{X}, \widehat{X})$. Using the fact that tensor is a multi-linear operator, and by the assumption that $\|\widehat{G}_3 - G_3\| \leq \delta_3$, we have:

$$
\begin{aligned}
\epsilon \equiv \|\widehat{G} - G\| &\leq \|\widehat{G}_3(\widehat{X}, \widehat{X}, \widehat{X}) - G_3(X, X, X)\|_F \\
&\leq \|G_3(\widehat{X}, \widehat{X}, \widehat{X}) - G_3(X, X, X)\| + \|\widehat{G}_3(\widehat{X}, \widehat{X}, \widehat{X}) - G_3(\widehat{X}, \widehat{X}, \widehat{X})\| \\
&\leq 3\|G_3(\Delta_X, X, X)\| + 3\|G_3(\Delta_X, \Delta_X, X)\| + \|G_3(\Delta_X, \Delta_X, \Delta_X)\| + \delta_3\|\widehat{X}\|^3 \\
&\leq 7\|G_3\|\|X\|^2\|\Delta_X\| + (\|X\| + \|\Delta_X\|)^3 \delta_3 \\
&\leq O\left(\frac{\|G_3\|}{\gamma_{min}^{2.5}} \delta_2 + \frac{1}{\gamma_{min}^{1.5}} \delta_3 \right).
\end{aligned}
$$

Note that by the assumption $\delta_2 \leq o(\frac{\gamma_{min}^{2.5}}{k\|G_3\|})$, $\delta_3 \leq o(\frac{\gamma_{min}^{1.5}}{k})$, we have $\epsilon \leq o(\frac{1}{k})$. Therefore we can apply Theorem D.2 to conclude that with probability at least $1 - e^{-n}$ (over the randomness of the randomized algorithm itself), the tensor power algorithm runs in time $\text{poly}(n, k, 1/\lambda_{min})$ and for some permutation π over $[k]$ it returns:

$$
\|\widehat{v}_{\pi(i)} - v_{\pi(i)}\| \leq \frac{8\epsilon}{\lambda_{min}}, \quad |\widehat{\lambda}_i - \lambda_i| \leq 5\epsilon, \quad \forall j \in [k].
$$

Finally, since we also have $5\epsilon \leq 1/2 \leq \lambda_{min}/2$ we can bound the estimation error of $\widehat{a}_i$ and $\widehat{\omega}_i$ as defined in (32) by:

$$
\begin{aligned}
\|\widehat{a}_{\pi(i)} - a_i\| &\leq 3(\|\Delta_X\|\lambda_{max} + \frac{1}{\gamma_{min}^{0.5}}\frac{8\epsilon}{\lambda_{min}}\lambda_{max} + \frac{1}{\gamma_{min}^{0.5}}5\epsilon) \\
&\leq \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2), 1/\omega_o)\delta_2 + \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2), 1/\omega_o)\delta_3, \\
\|\widehat{\omega}_i - \omega_i\| &\leq \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2))\delta_2 + \text{poly}(\|G_3\|, 1/\sigma_{min}(G_2))\delta_3.
\end{aligned}
$$

$\square$

Now we can apply Theorem D.2 to our case.

Lemma D.3. *Given $\widehat{Y}_4$, $\widehat{Y}_6$, $\widehat{U}$ and suppose that $\|\widehat{Y}_4 - \widetilde{Y}_4\|_F$, $\|\widehat{Y}_6 - \widetilde{Y}_6\|_F$ as well as $\|\widehat{U} - \widetilde{U}\|$ are bounded by some inverse $\text{poly}(n, k, 1/\omega_o, 1/\rho)\delta$. There exists an algorithm that with high probability, returns $\widehat{\Sigma}^{(i)}$'s and $\widehat{\omega}_i$'s such that for some permutation π over $[k]$, we have the distance $\|\widehat{\Sigma}^{(i)} - \widetilde{\Sigma}^{(i)}\|$ and $\|\widehat{\omega}_i - \widetilde{\omega}_i\|$ are bounded by δ. Moreover, the running time of the algorithm is upperbounded by $\text{poly}(n, k, 1/\omega_o, 1/\rho)$.*

Proof. (to Lemma D.3)

We apply Theorem D.2, and pick $G_2 = \widetilde{Y}_4$, $G_3 = \widetilde{Y}_6$. We only need to verify that $\|\widetilde{Y}_6\|$ and $1/\sigma_{min}(\widetilde{Y}_4)$ are polynomials of the relevant parameters. This is easy to see, since $\sigma_{min}(\widetilde{Y}_4) \geq \omega_o\sigma_{min}(\widetilde{\Sigma})^2$, and the matrix $\widetilde{\Sigma}$ is a perturbed rectangular matrix which by Lemma G.15 has $\sigma_{min}(\widetilde{\Sigma}) \geq \Omega(\rho\sqrt{n_2})$ with high probability.

Finally, given $\widehat{\sigma}^{(i)}$, and given the output of Step 2, i.e. $\widehat{U}$, with inverse polynomial accuracy, we can recover $\widehat{\Sigma}^{(i)} = \widehat{U}\widehat{\sigma}^{(i)}$ up to accuracy polynomial in the relevant parameters. $\square$

Input: Samples x_i from the mixture of Gaussians , number of components k.
Output: Set of parameters $\mathcal{G} = \{(\omega_i, \Sigma^{(i)}) : i \in [k]\}$.

Estimate M_4, M_6 using the samples.

$$M_4 = \frac{1}{N}\sum_{i=1}^{N} x_i \otimes^4, \quad M_6 = \frac{1}{N}\sum_{i=1}^{N} x_i \otimes^6 .$$

Let $s = 9\lceil\sqrt{n}\rceil$
(**Step 1 (a)** *Algorithm 1*)
$S_1 = \text{FindColumnSpan}(M_4, \{1, ..., s\})$,
$S_2 = \text{FindColumnSpan}(M_4, \{s+1, ..., 2s\})$.
(**Step 1 (b)** *Algorithm 2*)
$U_1 = \text{FindProjectedSigmaSpan}(M_4, \{1, ..., s\}, S_1)$,
$U_2 = \text{FindProjectedSigmaSpan}(M_4, \{s+1, ..., 2s\}, S_2)$.
(**Step 1 (c)** *Algorithm 3*)
$U = \text{MergeProjections}(S_1, U_1, S_2, U_2)$.
(**Step 2** *Algorithm 4*)
$(Y_4, Y_6) = \text{EstimateY}_4 Y_6(M_4, M_6, U)$.
(**Step 3** *Algorithm 5*)
$\mathcal{G} = \text{TensorDecomp}(Y_4, Y_6, U)$

Return: $\mathcal{G}$.

Algorithm 6: MainAlgorithm (Zero-mean case)

E   Proofs of Theorem 3.5

The results in all previous sections showed the correctness and robustness of each individual step for the algorithm for zero-mean case, In this section, we summarize those results to prove that the overall algorithm has polynomial time/sample complexity.

Lemma E.1 (Concentration of empirical moments). *Given N samples $x_1, \ldots, x_N$ drawn i.i.d. from the n-dimensional mixture of k Gaussians, if $N \geq n^7/\delta^2$, then with high probability, we have that for all $j_1, \ldots, j_6 \in [n]$:*

$$\left|[\widehat{M_4}]_{j_1,j_3,j_3,j_4} - [\widetilde{M_4}]_{j_1,j_3,j_3,j_4}\right| \leq \delta, \quad \left|[\widehat{M_6}]_{j_1,j_3,j_3,j_4,j_5,j_6} - [\widetilde{M_6}]_{j_1,j_3,j_3,j_4,j_5,j_6}\right| \leq \delta.$$

Proof. Let x denote the random vector of this mixture of Gaussians. We first truncate its tail probabilities to make all the entries ($[x]_j$ for $j \in [n]$) in the vector x be in the range $[-\sqrt{n}, \sqrt{n}]$. Apply union bound, we know that with high probability (at least $1 - O(e^{-n})$), for all indices $j_1, \ldots, j_6 \in [n]$, we have $\left|[x]_{j_1} \ldots [x]_{j_6}\right| \leq n^3$. Then we can apply Hoeffding's inequality to bound the empirical moments by:

$$\Pr\left[|\widehat{\mathbb{E}}[x_{j_1} \ldots x_{j_6}] - \mathbb{E}[x_{j_1} \ldots x_{j_6}]| \geq \delta\right] \leq \exp(-\frac{2\delta^2 N^2}{N(2n^3)^2}) + O(e^{-n}) \leq O(e^{-n}).$$

$\square$

Proof. (of Theorem 3.5)

44

We show that, to achieve ϵ accuracy in the output of Step 3 in the algorithm for the zero-mean case, the number of samples we need to estimate the moments M_4 and M_6 is bounded by a polynomial of relevant parameters, namely $\text{poly}(n, k, 1/\omega_o, 1/\epsilon, 1/\rho)$, and each step of the algorithm can be done in polynomial time.

We backtrack the input-output relations from Step 3 to Step 2 and to Step 1, and we show that the estimation error in the empirical moments and the inputs / outputs only *polynomially* propagate throughout the steps.

First note that we have shown that every steps fails with negligible probability ($O(e^{-n^C})$ for any absolute constant C). Then apply union bound, we have that the entire algorithm works correctly with high probability.

1. By Lemma D.3, in order to achieve ϵ accuracy in the final estimation of the mixing weights and the covariance matrices, we need to drive the input accuracy of Step 3 (also the output accuracy of Step 2) to be bounded by some inverse polynomial in $(n, 1/\epsilon, 1/\rho, 1/\omega_o)$, Also recall that this step has running time $\text{poly}(n, k, 1/\rho, 1/\omega_o)$.

2. Theorem C.1 and Theorem C.10 guarantee that with smoothed analysis $\sigma_{min}(\widetilde{H}_4)$ and $\sigma_{min}(\widetilde{H}_6)$ are lower bounded polynomially. Then by Lemma C.16, in order to have the output accuracy of Step 2 be bounded by inverse $\text{poly}(n, 1/\epsilon, 1/\rho, 1/\omega_o)$, we need to drive the input accuracy of Step 2 $(\widehat{U}, \widehat{M}_4)$ to be bounded by some other inverse polynomial. Step 2 involves solving linear systems of dimension $n_4 k_2$ and $n_6 k_3$, thus it running time is polynomial.

3. Lemma B.13 and B.14 guarantees that with smoothed analysis $\sigma_k(\widetilde{Q}_U)$ is lower bounded polynomially. Then by Lemma B.15, in order to have the output accuracy of Step 1 (c) $(\widehat{U})$ be bounded by inverse polynomial, we need to drive the input accuracy (output $\widehat{S}_i$ of Step 1 (a) and output $\widehat{U}_i$ of Step 1 (b)) to be bounded by some other inverse polynomial. Step 1 (c) involves multiplications and factorization of matrices of polynomial size, and thus the running time is also polynomial.

4. Lemma B.7 guarantees that with smoothed analysis $\sigma_k(\widetilde{Q}_{U_S})$ is lower bounded polynomially. Then by Lemma B.10, in order to have the output accuracy of Step 1 (b) $(\widehat{U}_S)$ be bounded by inverse polynomial, we need to drive the input accuracy (output $\widehat{S}_i$ of Step 1 (a)) to be bounded by some other inverse polynomial. Step 1 (b) involves multiplications and factorization of matrices of polynomial size, and thus the running time is also polynomial.

5. Lemma B.1 guarantees that with smoothed analysis $\sigma_k(\widetilde{Q}_S)$ is lower bounded by inverse polynomial. Then by Lemma B.3, in order to have the output accuracy of Step 1 (a) $(\widehat{S})$ be bounded by inverse polynomial, we need to drive the input accuracy (the moment estimation $\widehat{M}_4$) to be bounded by some other inverse polynomial. Step 1 (a) involves multiplications and factorization of matrices of polynomial size, and thus the running time is also polynomial.

6. Finally, by Lemma E.1, in order to have the accuracy of moment estimation $(\widehat{M}_4, \widehat{M}_6)$ be bounded by inverse polynomial, we need the number of samples N polynomial in all the relevant parameters, including k.

$\square$

F   General Case

In this section, we present the algorithm for learning mixture of Gaussians with general means. The algorithm generalizes the insights obtained from the algorithm for the zero-mean case. The

steps are very similar, and we will highlight the differences.

Input: Samples $\{x_i \in \mathbb{R}^n : i = 1, \ldots, N\}$ from the mixture of Gaussians, number of components k.
Output: Set of parameters $\mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)}) : i \in [k]\}$.
  Estimate M_3 M_4, M_6 using the samples

$$M_3 = \frac{1}{N} \sum_{i=1}^{N} x_i \otimes^3, \ M_4 = \frac{1}{N} \sum_{i=1}^{N} x_i \otimes^4, \ M_6 = \frac{1}{N} \sum_{i=1}^{N} x_6 \otimes^3$$

  Step 1 (a). *(This can be accomplished similar to Algorithm 1 FindColumnSpan)*
Let $\mathcal{H}_1 = \{1, \ldots, 12\sqrt{n}\}$, find $S_1 = \text{span}\{\widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)}_{[:,j]} : i \in [k], j \in \mathcal{H}_1\}$.
Let $\mathcal{H}_2 = \{12\sqrt{n} + 1, \ldots, 24\sqrt{n}\}$, find $S_2 = \text{span}\{\widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)}_{[:,j]} : i \in [k], j \in \mathcal{H}_2\}$.

  Step 1 (b) *(This can be accomplished similar to Algorithm 2 FindProjectedSigmaSpan)*
Find $U_1 = span\{\text{Proj}_{S_1^\perp} \widetilde{\Sigma}^{(i)} : i \in [k]\}$.
Find $U_2 = span\{\text{Proj}_{S_2^\perp} \widetilde{\Sigma}^{(i)} : i \in [k]\}$.

  Step 1 (c) *(This can be accomplished similar to Algorithm 3 MergeProjections)*
Merge U_1 and U_2 to get $Z = span\{\mu^{(i)} : i \in [k]\}$,
  $U' = span\{\text{vec}(\text{Proj}_{Z^\perp} \Sigma^{(i)}) : i \in [k]\}$, and $U_o = span\{\text{Proj}_{Z^\perp} \Sigma^{(i)} \text{Proj}_{Z^\perp} : i \in [k]\}$.

  Step 2
Project the samples to the subspace $Z^\perp$: $\text{Proj}_{Z^\perp} x = \{\text{Proj}_{Z^\perp} x_1, \ldots, \text{Proj}_{Z^\perp} x_N\}$.
Apply the algorithm for zero mean case to the projected samples,
let $\mathcal{G}_o = \{(\omega_i, \text{Proj}_{Z^\perp} \Sigma^{(i)} \text{Proj}_{Z^\perp}) : i \in [k]\} = \text{MainAlgorithm (Zero-mean case)}(\text{Proj}_{Z^\perp} x)$.

  Step 3
Let $T = \left[\text{vec}(\text{Proj}_{Z^\perp} \Sigma^{(i)} \text{Proj}_{Z^\perp}) : i \in [k] \right]^{\dagger\top} \in \mathbb{R}^{n^2 \times k}$,
and let $T^{(i)}$ for $i \in [k]$ denote the columns of T.
Let $M_{3(1)} \in \mathbb{R}^{n \times n^2}$ be the matricization of M_3 along the first dimension.
Let $\mu^{(i)} = M_{3(1)} T^{(i)} / \omega_i$ for $i \in [k]$ and let $\mu = [\mu^{(i)} : i \in [k]]$.

  Step 4
  Let $M_4' = M_4 + 2 \sum_{i=1}^{k} \omega_i \mu^{(i)} \otimes^4$.
  Find the span $S = span\{\text{vec}(\widetilde{\Sigma}^{(i)}) + \widetilde{\mu}^{(i)} \odot \widetilde{\mu}^{(i)} : i \in [k]\}$.
(This can be achieved by treating M_4' as the 4-th moments of a mixture of zero-mean Gaussians, and apply Step 1 in the algorithm for zero-mean case to find the span of the covariance matrices, and let S denote the result.)
  Let $\Sigma = [\text{vec}(\Sigma^{(i)}) : i \in [k]] = (\text{Proj}_S U' - \mu \odot \mu)$.

Return: $\mathcal{G} = \{(\omega_i, \mu^{(i)}, \Sigma^{(i)}) : i \in [k]\}$.

Algorithm 7: MainAlgorithm (General Case)

Step 1. Span finding In this step, we find the following two subspaces:

$$\widetilde{Z} = span\{\widetilde{\mu}^{(i)} : i \in [k]\}, \quad \widetilde{\Sigma}_o = span\{\text{Proj}_{\widetilde{Z}^\perp}\widetilde{\Sigma}^{(i)}\text{Proj}_{\widetilde{Z}^\perp}\}.$$

This is very similar to Step 1 in the algorithm for the zero-mean case, and can be achieved in three small steps:

1. Step 1 (a). For a subset $\mathcal{H}$ of size $12\sqrt{n}$, find the span $\mathcal{S}$ of the mean vectors and a subset of columns of the covariance matrices:

$$\mathcal{S} = \text{span}\{\widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)}_{[:,j]} : i \in [k], j \in \mathcal{H}\}.$$

2. Step 1 (b). Find the span of covariance matrices projected to the subspace $S^\perp$:

$$\mathcal{U}_S = span\{\text{Proj}_{S^\perp}\widetilde{\Sigma}^{(i)} : i \in [k]\}.$$

3. Step 1 (c). Run 1(a) and 1(b) on two disjoint subsets $\mathcal{H}_1$ and $\mathcal{H}_2$. Merge the two spans U_1 and U_2 to get $\widetilde{Z}$ and $span\{\text{Proj}_{\widetilde{Z}^\perp}\widetilde{\Sigma}^{(i)} : i \in [k]\}$.

Next, we discuss each small step and compare it with the similar analysis of the algorithm for the zero-mean case.

Step 1 (a). Find the span $\mathcal{S}$ of the means and a subset of the columns of the covariance matrices Similar to Step 1 (a) for the zero-mean case, in this step we want to find a subspace $\mathcal{S}$ which contains the span of a subset of columns of $\widetilde{\Sigma}^{(i)}$'s. However, with the mean vector $\widetilde{\mu}^{(i)}$'s appearing in the moments, the subspace we find also contains the span of all the mean vectors. In particular, for a subset $\mathcal{H} \in [n]$ with $|\mathcal{H}| = \sqrt{n}$, we aim to find the following subspace:

$$\mathcal{S} = span\{\widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)}_{[:,j]} : i \in [k], j \in \mathcal{H}\}. \tag{33}$$

Similar to Claim 5.1 for the zero-mean case, the key observation for finding the subspace is the structure of the one-dimensional slices of the 4-th order moments for the general case:

Claim F.1. *For any indices $j_1, j_2, j_3 \in [n]$, the one-dimensional slices of $\widetilde{M}_4$ are given by:*

$$\widetilde{M}_4(e_{j_1}, e_{j_2}, e_{j_3}, I) = \sum_{i=1}^n \widetilde{\omega}_i \left(\widetilde{\mu}^{(i)}_{j_1}\widetilde{\mu}^{(i)}_{j_2}\widetilde{\mu}^{(i)}_{j_3}\widetilde{\mu}^{(i)} + \sum_{\pi \in \left\{\begin{smallmatrix}(j_1,j_2,j_3),\\(j_2,j_3,j_1),\\(j_3,j_1,j_2)\end{smallmatrix}\right\}} \widetilde{\Sigma}^{(i)}_{\pi_1,\pi_2}\widetilde{\Sigma}^{(i)}_{[:,\pi_3]} + \widetilde{\mu}^{(i)}_{\pi_1}\widetilde{\mu}^{(i)}_{\pi_2}\widetilde{\Sigma}^{(i)}_{[:,\pi_3]} + \widetilde{\Sigma}^{(i)}_{\pi_1,\pi_2}\widetilde{\mu}^{(i)}_{\pi_3}\widetilde{\mu}^{(i)} \right)$$

$$\tag{34}$$

Note that if we pick the indices $j_1, j_2, j_3 \in \mathcal{H}$, all such one-dimensional slice of $\widetilde{M}_4$ lie in the subspace $\mathcal{S}$. We again evenly partition the set $\mathcal{H}$ into three disjoint subset $\mathcal{H}^{(i)}$ and take $j_i \in \mathcal{H}^{(i)}$ for $i = 1, 2, 3$. Define the matrix $\widetilde{Q}_S \in \mathbb{R}^{n \times (|\mathcal{H}|/3)^3}$ as in (12) whose columns are the one-dimensional slices of $\widetilde{M}_4$:

$$\widetilde{Q}_S = \left[\left[[\widetilde{M}_4(e_{j_1}, e_{j_2}, e_{j_3}, I) : j_3 \in \mathcal{H}^{(3)}] : j_2 \in \mathcal{H}^{(2)} \right] : j_1 \in \mathcal{H}^{(1)} \right] \in \mathbb{R}^{n \times (|\mathcal{H}|/3)^3}. \tag{35}$$

The proof of this step is similar to the Lemmas B.1 (for smoothed analysis) and B.3 (for stability analysis). The main difference is that in the matrix $\widetilde{B}$ defined in the structural Claim B.2, there is

now another block $\widetilde{B}^{(0)}$ with k columns that corresponds to the $\widetilde{\mu}^{(i)}$ directions, which we can again handle with Lemma G.12.

Lemma F.2 shows the deterministic conditions for Step 1 (a) to correctly identify the subspace $\mathcal{S}$ from the columns of $\widetilde{Q}_S$, and uses smoothed analysis to show that the conditions hold with high probability.

Lemma F.2 (Correctness). *Given $\widetilde{M_4}$ of a general mixture of Gaussians , for any subset $\mathcal{H} \in [n]$ and $|\mathcal{H}| = c_2 k$ with the constant $c_2 > 9$, let $\widetilde{Q}_S$ be the matrix defined as in (35). The columns of $\widetilde{Q}_S$ give the desired span S defined in (33) if the matrix $\widetilde{Q}_S$ achieves the maximal column rank $k + k|\mathcal{H}|$. With probability (over the ρ-perturbation) at least $1 - C\epsilon^{0.5n}$ for some constant C, the $k(1 + |\mathcal{H}|)$-th singular value of $\widetilde{Q}_S$ is bounded below by:*

$$\sigma_{k(1+|\mathcal{H}|)}(\widetilde{Q}_S) \geq \rho\epsilon\sqrt{n}.$$

The proof idea is similar to that of Lemma B.1. We construct a basis $\widetilde{P}_S \in \mathbb{R}^{n \times (k+k|\mathcal{H}|)}$ for the subspace $\mathcal{S}$ as follows.

$$\widetilde{P}_S = \left[\left[\widetilde{\mu}^{(i)} : i \in [k]\right], \left[[\widetilde{\Sigma}^{(i)}_{[:,j]} : i \in [k]] : j \in \mathcal{H}^{(l)}\right] : l = 1, 2, 3\right] = \left[\widetilde{\mu}, \ \widetilde{\Sigma}_{[:,\mathcal{H}^{(1)}]}, \widetilde{\Sigma}_{[:,\mathcal{H}^{(2)}]}, \widetilde{\Sigma}_{[:,\mathcal{H}^{(3)}]}\right]. \tag{36}$$

Note that the dimension of the subspace $\mathcal{S}$ is at most $k(|\mathcal{H}| + 1) < n/3$. Then we show by the Claim about the moment structure that the matrix $\widetilde{Q}_S$ can be written as a product of $\widetilde{P}_S$ and some coefficient matrix $\widetilde{B}_S$. Then we bound the smallest singular value of the two matrices $\widetilde{P}_S$ and $\widetilde{B}_S$ via smoothed analysis separately. The coefficient matrix $\widetilde{B}_S$ is slightly different than that in the zero-mean case, but has similar block-diagonal structure properties.

The detailed proof is provided below.

Proof. (of Proposition F.2)

Similar to structural property in Claim B.2 for the zero-mean case, we can write the matrix $\widetilde{Q}_S$ in a product form:

$$\widetilde{Q}_S = \widetilde{P}_S \left(D_{\widetilde{\omega}} \otimes_{kr} I_{|\mathcal{H}|}\right) (\widetilde{B}_S)^{\top}.$$

We will bound the smallest singular value for each of the factor, and apply union bound to conclude the lower bound of $\sigma_{k(1+|\mathcal{H}|)}(\widetilde{Q}_S)$.

The matrix $\widetilde{P}_S \in \mathbb{R}^{n \times (k+k|\mathcal{H}|)}$ is defined in (36). Restricting to the rows corresponding to $[n]\backslash\mathcal{H}$, we can use Lemma G.16 to argue that $\sigma_{k(1+|\mathcal{H}|)} \geq \epsilon\rho\sqrt{n}$ with probability at least $1 - (C\epsilon)^{0.25n}$.

In order to lower bound $\sigma_{min}(\widetilde{B}_S)$, we first analyze the structure of this coefficient matrix. The matrix $\widetilde{B}_S$ has the following block structure:

$$\widetilde{B}_S = \left[\widetilde{B}^{(0)}, \widetilde{B}^{(1)}, \widetilde{B}^{(2)}, \widetilde{B}^{(3)}\right].$$

The first block $\widetilde{B}^{(0)} \in \mathbb{R}^{(|\mathcal{H}|/3)^3 \times k}$ is a summation of four matrices $\widetilde{B}^{(0)}_i$ for $i = 0, 1, 2, 3$, where $\widetilde{B}^{(0)}_0 = \widetilde{\mu}_{\mathcal{H}^{(3)}} \odot \widetilde{\mu}_{\mathcal{H}^{(2)}} \odot \widetilde{\mu}_{\mathcal{H}^{(1)}}$, and $\widetilde{B}^{(0)}_1 = \widetilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(2)}} \odot \widetilde{\mu}_{\mathcal{H}^{(1)}}$. With some fixed and known row permutation $\pi^{(2)}$ and $\pi^{(3)}$, the other two matrix blocks $\widetilde{B}^{(0)}_2$ and $\widetilde{B}^{(0)}_3$ are equal to $\widetilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(1)}} \odot \widetilde{\mu}_{\mathcal{H}^{(2)}}$ and $\widetilde{\Sigma}_{\mathcal{H}^{(2)}, \mathcal{H}^{(1)}} \odot \widetilde{\mu}_{\mathcal{H}^{(3)}}$, separately.

The block $\widetilde{B}^{(1)} \in \mathbb{R}^{(|\mathcal{H}|/3)^3 \times k|\mathcal{H}|/3}$ is block diagonal with the identical block $\widetilde{\Sigma}_{\mathcal{H}^{(3)}, \mathcal{H}^{(2)}} + \widetilde{\mu}_{\mathcal{H}^{(3)}} \odot \widetilde{\mu}_{\mathcal{H}^{(2)}}$. Similarly, with the row permutation $\pi^{(2)}$, $\pi^{(3)}$, the other two matrix blocks $\widetilde{B}^{(2)}, \widetilde{B}^{(3)}$ are

48

equal to the block diagonal matrices with the identical block $(\widetilde{\Sigma}_{\mathcal{H}^{(3)},\mathcal{H}^{(1)}} + \widetilde{\mu}_{\mathcal{H}^{(3)}} \odot \widetilde{\mu}_{\mathcal{H}^{(1)}})$ and $(\widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(1)}} + \widetilde{\mu}_{\mathcal{H}^{(2)}} \odot \widetilde{\mu}_{\mathcal{H}^{(1)}})$ respectively.

Note that we can write the block $\widetilde{B}^{(0)}$ as:

$$
\begin{aligned}
\widetilde{B}^{(0)} =& (\widetilde{\mu}_{\mathcal{H}^{(3)}} \odot \widetilde{\mu}_{\mathcal{H}^{(2)}} + \widetilde{\Sigma}_{\mathcal{H}^{(3)},\mathcal{H}^{(2)}}) \odot \widetilde{\mu}_{\mathcal{H}^{(1)}} + (\pi^{(2)})^{-1}(\widetilde{\mu}_{\mathcal{H}^{(3)}} \odot \widetilde{\mu}_{\mathcal{H}^{(1)}} + \widetilde{\Sigma}_{\mathcal{H}^{(3)},\mathcal{H}^{(1)}}) \odot \widetilde{\mu}_{\mathcal{H}^{(2)}} \\
& + (\pi^{(3)})^{-1}(\widetilde{\mu}_{\mathcal{H}^{(2)}} \odot \widetilde{\mu}_{\mathcal{H}^{(1)}} + \widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(1)}}) \odot \widetilde{\mu}_{\mathcal{H}^{(3)}} - 2\widetilde{\mu}_{\mathcal{H}^{(3)}} \odot \widetilde{\mu}_{\mathcal{H}^{(2)}} \odot \widetilde{\mu}_{\mathcal{H}^{(1)}},
\end{aligned}
$$

where it is easy to see the first summand $(\widetilde{\mu}_{\mathcal{H}^{(3)}} \odot \widetilde{\mu}_{\mathcal{H}^{(2)}} + \widetilde{\Sigma}_{\mathcal{H}^{(3)},\mathcal{H}^{(2)}}) \odot \widetilde{\mu}_{\mathcal{H}^{(1)}}$ is a linear combination of the columns of the block diagonal matrix $\widetilde{B}^{(1)}$, and similarly the second and third summands are linear combinations of the columns of $\widetilde{B}^{(2)}$ and $\widetilde{B}^{(3)}$, and the last summand is simply $-2\widetilde{B}_0^{(0)}$. Therefore for some absolute constant C (the smallest singular value corresponding to the linear transformation) we have that:

$$
\sigma_{min}(\widetilde{B}_S) \geq C\sigma_{min}\left(\left[\widetilde{B}_0^{(0)}, \widetilde{B}^{(1)}, \widetilde{B}^{(2)}, \widetilde{B}^{(3)}\right]\right)
$$

Note that $\widetilde{B}_0^{(0)} = \widetilde{\mu}_{\mathcal{H}^{(3)}} \odot \widetilde{\mu}_{\mathcal{H}^{(2)}} \odot \widetilde{\mu}_{\mathcal{H}^{(1)}}$ only depends on the randomness over the mean vectors. Note that the Khatri-Rao product is a submatrix of the Kronecker product, therefore for tall matrices Q_1 and Q_2, we have that $\sigma_{min}(Q_1 \odot Q_2) \leq \sigma_{min}(Q_1 \otimes_{kr} Q_2) = \sigma_{min}(Q_1)\sigma_{min}(Q_2)$. In particular, we can bound the smallest singular value of $\widetilde{B}_0^{(0)}$ with high probability (at least $1-C\epsilon^{0.5n}$) as follows:

$$
\sigma_k(\widetilde{B}_0^{(0)}) \geq \sigma_k(\widetilde{\mu}_{\mathcal{H}^{(3)}})\sigma_k(\widetilde{\mu}_{\mathcal{H}^{(2)}})\sigma_k(\widetilde{\mu}_{\mathcal{H}^{(1)}}) \geq (\rho\epsilon\sqrt{n})^3.
$$

Then condition on the value of the means, we further exploit the randomness over the covariance matrices to lower bound $\sigma_{k|\mathcal{H}|}\left(\text{Proj}_{\widetilde{B}_0^{(0)\perp}}[\widetilde{B}^{(1)}, \widetilde{B}^{(2)}, \widetilde{B}^{(3)}]\right)$. It is almost the same as the argument of the proof for Proposition B.1. For example, compared to (18) we have the following inequality instead:

$$
\sigma_k\left(\text{Proj}_{([\widetilde{B}^{(0)}, \widetilde{B}^{(2)}, \widetilde{B}^{(3)}]_{\{j\}\times\mathcal{H}^{(2)}\times\mathcal{H}^{(3)}})^\perp}\text{Proj}_{(\Sigma_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}} + \widetilde{\mu}_{\mathcal{H}^{(2)}} \odot \widetilde{\mu}_{\mathcal{H}^{(3)}})^\perp}(\widetilde{\Sigma}_{\mathcal{H}^{(2)},\mathcal{H}^{(3)}} + \widetilde{\mu}_{\mathcal{H}^{(2)}} \odot \widetilde{\mu}_{\mathcal{H}^{(3)}})\right) \geq \epsilon\rho\sqrt{n},
$$

and note that any block in $\widetilde{B}^{(0)}$ is independent of the randomness of covariance matrices, and we have $(|\mathcal{H}|/3)^2 - k - 2k|\mathcal{H}|/3 \geq 2k$. Similar modifications apply to the inequalities in (20),(21).

Finally by the argument of Lemma G.12 we can bound $\sigma_{min}(\widetilde{B}_S)$ with probability at least $1 - C\epsilon^{0.5n}$ (over the randomness of both the perturbed means and covariance matrices):

$$
\sigma_{min}(\widetilde{B}_S) \geq \min\{(\rho\epsilon\sqrt{n})^3, \epsilon\rho\sqrt{n}\} = \epsilon\rho\sqrt{n},
$$

as we assume ρ to be small perturbation and $\rho\epsilon\sqrt{n} < 1$.

$\square$

Step 1 (b). Find the projected span of covariance matrices Given the subspace $\mathcal{S} = span\{\widetilde{\mu}^{(i)}, \widetilde{\Sigma}_{[:,\mathcal{H}]}^{(i)} : i \in [k]\}$ obtained from Step 1 (a), Step 1(b) finds the span of the covariance matrices with the columns projected to $S^\perp$, namely:

$$
\mathcal{U}_S = span\{\text{Proj}_{S^\perp}\widetilde{\Sigma}^{(i)} : i \in [k]\}.
$$

This is in parallel with Step 1 (b) for the zero-mean case, and we rely on the structure of the two-dimensional slices of $\widetilde{M}_4$ to find the span of the projected covariance matrices. Similar to Claim B.6 for the zero-mean case, the following claim shows how the structure of the two-dimensional slices is related to the desired span.

Claim F.3. *For a mixture of general Gaussians, the two-dimensional slices of $\widetilde{M}_4$ are given by:*

$$\widetilde{M}_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I) = \sum_{i=1}^{k} \widetilde{\omega}_i \Big((\widetilde{\Sigma}^{(i)}_{j_1,j_2} + \widetilde{\mu}^{(i)}_{j_1}(\widetilde{\mu}^{(i)}_{j_2})^\top)(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top)$$

$$+ \widetilde{\mu}^{(i)}_{j_1}(\widetilde{\mu}^{(i)}(\widetilde{\Sigma}^{(i)}_{[:,j_2]})^\top + \widetilde{\Sigma}^{(i)}_{[:,j_2]}(\widetilde{\mu}^{(i)})^\top) + \widetilde{\mu}^{(i)}_{j_2}(\widetilde{\mu}^{(i)}(\widetilde{\Sigma}^{(i)}_{[:,j_1]})^\top + \widetilde{\Sigma}^{(i)}_{[:,j_1]}(\widetilde{\mu}^{(i)})^\top)$$

$$+ \widetilde{\Sigma}^{(i)}_{[:,j_1]}(\widetilde{\Sigma}^{(i)}_{[:,j_2]})^\top + \widetilde{\Sigma}^{(i)}_{[:,j_2]}(\widetilde{\Sigma}^{(i)}_{[:,j_1]})^\top \Big), \quad \forall j_1, j_2 \in [n].$$

Note that given the set of indices $\mathcal{H}$ we chose in Step 1 (a) and the subspace S, if we pick the indices $j_1, j_2 \in \mathcal{H}$, project the two-dimensional slice to $S^\perp$, all the rank one terms in the sum are eliminated and the projected slice lies in the desired span $\mathcal{U}_S$:

$$\mathrm{Proj}_{S^\perp} \widetilde{M}_4(\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, I, I) = \sum_{i=1}^{k} \widetilde{\omega}_i (\widetilde{\Sigma}^{(i)}_{j_1,j_2} + \widetilde{\mu}^{(i)}_{j_1}(\widetilde{\mu}^{(i)}_{j_2})^\top) \mathrm{Proj}_{S^\perp} \widetilde{\Sigma}^{(i)}, \quad \forall j_1, j_2 \in \mathcal{H}.$$

Applying the same argument as in Lemma B.7 for the zero-mean case, we can show that with high probability over the perturbation, all the projected slices span the subspace $\mathcal{U}_S$.

Step 1 (c). Merge the two projections of covariance matrices Pick two disjoint index set $\mathcal{H}_1$ and $\mathcal{H}_2$ and repeat the previous two steps 1 (a) and 1 (b), we can obtain the two spans U_1 and U_2, corresponding to the subspace of the covariance matrices projected to $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively.

In this step, we apply similar techniques as in Step 1 (c) for the zero-mean case to merge the two spans U_1 and U_2: we first use the overlapping part of the two projections $\mathrm{Proj}_{S_1^\perp}$ and $\mathrm{Proj}_{S_2^\perp}$ to align the basis of U_1 and U_2, then merge the two spans using the same basis.

Note that for the general case, by definition the span of the mean vectors $\widetilde{Z}$ lie in both subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, therefore we have $\mathcal{S}_1^\perp \subset \widetilde{Z}^\perp$ and $\mathcal{S}_2^\perp \subset \widetilde{Z}^\perp$. We can show that $\mathcal{S}_1^\perp \cup \mathcal{S}_2^\perp = \widetilde{Z}^\perp$ by lower bounding $\sigma_{n-k}([\mathrm{Proj}_{\mathcal{S}_1^\perp}, \mathrm{Proj}_{\mathcal{S}_2^\perp}])$ with high probability, similar to that in (28). This gives us the span of the mean vectors $\widetilde{Z}$.

Moreover, in the general case, from merging U_1 and U_2 we are only able to find the span of covariance matrices projected to the subspace $\widetilde{Z}^\perp$. In particular, we can follow Lemma B.11 and Lemma B.15 in Step 1 (c) for the zero-mean case to show that for the general case, we can merge U_1 and U_2 to obtain the span $span\{\mathrm{Proj}_{\widetilde{Z}^\perp} \widetilde{\Sigma}^{(i)} : i \in [k]\}$. By further projecting the span to $\widetilde{Z}^\perp$ from the right side, we can also obtain $\widetilde{\Sigma}_o = span\{\mathrm{Proj}_{\widetilde{Z}^\perp} \widetilde{\Sigma}^{(i)} \mathrm{Proj}_{\widetilde{Z}^\perp} : i \in [k]\}$.

Step 2. Find the covariance matrices in the subspace orthogonal to the means Given the subspace $\widetilde{Z}$ and $\widetilde{\Sigma}_o = span\{\mathrm{Proj}_{\widetilde{Z}^\perp} \widetilde{\Sigma}^{(i)} \mathrm{Proj}_{\widetilde{Z}^\perp} : i \in [k]\}$ obtained from Step 1, Step 2 applies the zero-mean case algorithm to find the covariance matrices projected to the subspace $\widetilde{Z}^\perp$, i.e., $\mathrm{Proj}_{\widetilde{Z}^\perp} \widetilde{\Sigma}^{(i)} \mathrm{Proj}_{\widetilde{Z}^\perp}$'s, as well as find the mixing weights $\widetilde{\omega}_i$'s.

This follows the same arguments as in Step 2 and Step 3 for the zero mean case. Consider projecting all the samples to $\widetilde{Z}^\perp$, the subspace orthogonal to all the means. In this subspace, the samples are like from a mixture of zero-mean Gaussians with the projected covariance matrices, and the 4-th and 6-th order moment are given by $\widetilde{M}_4(\mathrm{Proj}_{\widetilde{Z}^\perp}, \mathrm{Proj}_{\widetilde{Z}^\perp}, \mathrm{Proj}_{\widetilde{Z}^\perp}, \mathrm{Proj}_{\widetilde{Z}^\perp})$ and $\widetilde{M}_6(\mathrm{Proj}_{\widetilde{Z}^\perp}, \mathrm{Proj}_{\widetilde{Z}^\perp}, \mathrm{Proj}_{\widetilde{Z}^\perp}, \mathrm{Proj}_{\widetilde{Z}^\perp}, \mathrm{Proj}_{\widetilde{Z}^\perp}, \mathrm{Proj}_{\widetilde{Z}^\perp})$. Since $\widetilde{Z}$ is of dimension k, the dimension of the zero-mean Gaussian in the projected space is at least $n - k = O(n)$.

Note that the subspace $\widetilde{Z}^\perp$ only depends on the randomness of the means, and random perturbation on the covariance matrices is independent of that of $\widetilde{\mu}$. The smoothed analysis for the moment unfolding in Step 2 and tensor decomposition in Step 3 for the zero-mean case, which only depend on the randomness of the covariance matrices, still go through in the projected space.

50

Step 3. Find the means This step finds the mean vectors based on the outputs of the previous steps. The key observation for this step is about the structure of the 3-rd order moments in the following claim:

Claim F.4. *Let the matrix $\widetilde{M}_{3(1)} \in \mathbb{R}^{n \times n^2}$ be the matricization of $\widetilde{M}_3$ along the first dimension. The j-th row of $\widetilde{M}_{3(1)}$ is given by:*

$$[\widetilde{M}_{3(1)}]_{[j,:]} = \left[\left[\mathbb{E}[x_j x_{j_1} x_{j_2}] : j_1 \in [n] \right] : j_2 \in [n] \right]$$

$$= \sum_{i=1}^{k} \widetilde{\omega}_i \left(\widetilde{\mu}_j^{(i)} vec(\widetilde{\Sigma}^{(i)}) + \widetilde{\mu}_j^{(i)} \widetilde{\mu}^{(i)} \odot \widetilde{\mu}^{(i)} + \widetilde{\Sigma}_{[:,j]}^{(i)} \odot \widetilde{\mu}^{(i)} + \widetilde{\mu}^{(i)} \odot \widetilde{\Sigma}_{[:,j]}^{(i)} \right)^{\top} \qquad (37)$$

The following lemma shows how to extract the means $\widetilde{\mu}^{(i)}$'s from $\widetilde{M}_{3(1)}$ using the information of the covariance matrices projected to the subspace orthogonal to the means, i.e. $\widetilde{\Sigma}_o$, and the mixing weights $\widetilde{\omega}_i$'s.

Lemma F.5. *Given the mixing weights $\widetilde{\omega}_i$'s and the projected covariances $\widetilde{\Sigma}_o^{(i)}$'s, define the matrix $\widetilde{T} \in \mathbb{R}^{n^2 \times k}$ to be the pseudo-inverse of $\widetilde{\Sigma}_o$:*

$$\widetilde{T} = \left[vec(\widetilde{\Sigma}_o^{(i)}) : i \in [k] \right]^{\dagger\top}.$$

The mean $\widetilde{\mu}^{(i)}$ of the i-th component can be obtained by:

$$\widetilde{\mu}^{(i)} = \frac{1}{\widetilde{\omega}_i} \widetilde{M}_{3(1)} \widetilde{T}_{[:,i]}.$$

This step correctly finds the means if the $\widetilde{\Sigma}_o$ is full rank with good condition number, and this holds with high probability over the perturbation.

Proof. (of Lemma F.5)

The basic idea is that since $\widetilde{\Sigma}_o$ lies in the span of $\widetilde{P} = \mathrm{Proj}_{\widetilde{Z}^\perp} \otimes_{kr} \mathrm{Proj}_{\widetilde{Z}^\perp}$, and the last three summands in the parenthesis in (37) all lie in $span\{I_n \otimes_{kr} \mathrm{Proj}_{\widetilde{Z}}, \ \mathrm{Proj}_{\widetilde{Z}} \otimes_{kr} I_n\} = span\{\widetilde{P}^\perp\}$. Therefore hitting the matrix $\widetilde{M}_{3(1)}$ with $\widetilde{\Sigma}_o^\dagger$ from the right will eliminate those summands and pull out only the mean vectors.

Recall that the columns of the matrix $\widetilde{\Sigma}_o$ are $\mathrm{vec}(\mathrm{Proj}_{\widetilde{Z}^\perp} \widetilde{\Sigma}^{(i)} \mathrm{Proj}_{\widetilde{Z}^\perp}) = \widetilde{P}\mathrm{vec}(\widetilde{\Sigma}^{(i)})$'s, and the columns of $\widetilde{\Sigma}$ are $\mathrm{vec}(\widetilde{\Sigma}^{(i)})$'s.

Note that $\widetilde{T} = (\widetilde{P}\widetilde{\Sigma})^{\dagger\top} = \widetilde{P}\widetilde{\Sigma}^{\dagger\top}$, and the columns of $\widetilde{T}$ lie in $span\{\widetilde{P}\}$. Also note that for all $i,j \in [k]$ the vectors $\widetilde{\mu}^{(i)} \odot \widetilde{\mu}^{(i)}$, $\widetilde{\Sigma}_{[:,j]}^{(i)} \odot \widetilde{\mu}^{(i)}$ and $\widetilde{\mu}^{(i)} \odot \widetilde{\Sigma}_{[:,j]}^{(i)}$ all lie in the subspace $span\{I_n \otimes_{kr} \mathrm{Proj}_{\widetilde{Z}}, \ \mathrm{Proj}_{\widetilde{Z}} \otimes_{kr} I_n\} = span\{\widetilde{P}^\perp\}$. Therefore these terms will be eliminated if we multiply the columns of $\widetilde{T}$ to the right of $\widetilde{M}_{3(1)}$. For the first term $\widetilde{\mu}_j^{(i)}\mathrm{vec}(\widetilde{\Sigma}^{(i)})$, since $\mathrm{vec}(\widetilde{\Sigma}^{(j)})^\top \widetilde{T}_{[:,i]} = (\widetilde{P}\mathrm{vec}(\widetilde{\Sigma}^{(j)}))^\top \widetilde{T}_{[:,i]} = 1_{[i=j]}$. Therefore, we have $\widetilde{M}_{3(1)}\widetilde{T}_{[:,i]} = \widetilde{\omega}_i \widetilde{\mu}^{(i)}$.

The smoothed analysis for the correctness of this step is easy. We only need to show that both $\widetilde{\Sigma}_o$ and $\widetilde{\Sigma}$ robustly have full column rank with high probability over perturbation of the covariance matrices, and thus the pseudo-inverse $\widetilde{T}$ is well defined. This follows from Lemma G.15.

Finally, the stability analysis for this step is also straightforward using the perturbation bound for pseudo-inverse in Theorem G.7. □

Step 4. Find the unprojected covariance matrices Note that by definition $\widetilde{Z} = span\{\widetilde{\mu}^{(i)} : i \in [k]\}$, the projected covariance $\text{Proj}_{\widetilde{Z}^\perp}(\widetilde{\Sigma}^{(i)})$ we obtained in Step 2 is also equal to $\text{Proj}_{\widetilde{Z}^\perp}(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top)$. In Step 4 we try to recover the missing part of the covariance matrices in the subspace $\widetilde{Z}$. Note that since we have also obtained the means in Step 3, it is equivalent to finding $(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top)$ for all i. We will show that if we can find the $span\{(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top) : i \in [k]\}$, the projected vector $\text{Proj}_{\widetilde{Z}^\perp}(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top)$ can be used as anchor to pin down the unprojected vector.

They key observation for finding the span of $span\{(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top) : i \in [k]\}$ is to first construct a 4-th order tensor $\widetilde{M}'_4$ which corresponds to the 4-th moment of a mixture of zero-mean Gaussians with covariance matrices $(\widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top)$, and then follow Step 1 in the algorithm for zero-mean case to find the span of the covariance matrices for this new mixture of Gaussians.

The next lemma shows how to construct such 4-th order tensor:

Lemma F.6. *Given the 4-th moment $\widetilde{M}_4$ for a mixture of Gaussians with parameters $\{\widetilde{\omega}_i, \widetilde{\mu}^{(i)}, \widetilde{\Sigma}^{(i)}\}$, define the 4-th order tensor $\widetilde{M}'_4$ to be:*

$$\widetilde{M}'_4 = \widetilde{M}_4 + 2\sum_{i=1}^{k}\widetilde{\omega}_i\widetilde{\mu}^{(i)}\otimes^4,$$

then $\widetilde{M}'_4$ is equal to the 4-th moment of a mixture Gaussians with parameters $\{\widetilde{\omega}_i, 0, \widetilde{\Sigma}^{(i)} + \widetilde{\mu}^{(i)}(\widetilde{\mu}^{(i)})^\top\}$.

The proof follows directly from Isserlis' Theorem. Therefore we can repeat Step 1 in the zero-mean case here to find the span of the space $\{\text{vec}(\widetilde{\Sigma}^{(i)}) + \widetilde{\mu}^{(i)} \odot \widetilde{\mu}^{(i)} : i \in [k]\}$. Since we also know the projection of $\widetilde{\Sigma}^{(i)}$'s in a large subspace (in the subspace $\text{Proj}_{\widetilde{Z}^\perp} \otimes_{kr} \text{Proj}_{\widetilde{Z}^\perp}$ obtained from Step 2), we can easily recover $\widetilde{\Sigma}^{(i)}$'s:

Lemma F.7. *For any matrix $U \in \mathbb{R}^{d \times k}$ and any subspace P, given $P^\top U$ and the span S of columns of U, the matrix U can be computed as*

$$U = S(P^\top S)^\dagger(P^\top U).$$

Further, this procedure is stable if $\sigma_{min}(P^\top S)$ is lower bounded.

Proof. This is a special case of the Step 1 (c) where we merge two projections of an unknown subspace.

The span S is equal to UV for some unknown matrix V. We can compute $V = (P^\top U)^\dagger P^\top S$, and hence $U = SV^{-1} = S(P^\top S)^\dagger(P^\top U)$. The stability analysis is similar (and simpler than) Lemma B.11. $\qquad\square$

We will apply this lemma to where the subspace P is $\text{Proj}_{\widetilde{Z}^\perp} \otimes_{kr} \text{Proj}_{\widetilde{Z}^\perp}$. Since the perturbation of the means and the covariance matrices are independent, we can lower bound the smallest singular value of $P^\top S$.

F.1 Proof Sketch of the Main Theorem 3.4

The proof follows the same strategy as Theorem 3.5. First we apply the union bound to all the smoothed analysis lemmas, this will ensure the matrices we are inverting all have good condition number, and the whole algorithm is robust to noise.

Then in order to get the desired accuracy ϵ, we need to guarantee inverse polynomial accuracy in different steps (through the stability lemmas). The flow of the algorithm is illustrated in Figure 5. In the end all the requirements becomes a inverse polynomial accuracy requirement on $\widehat{M}_4$ and $\widehat{M}_6$, which we obtain by Lemma E.1.
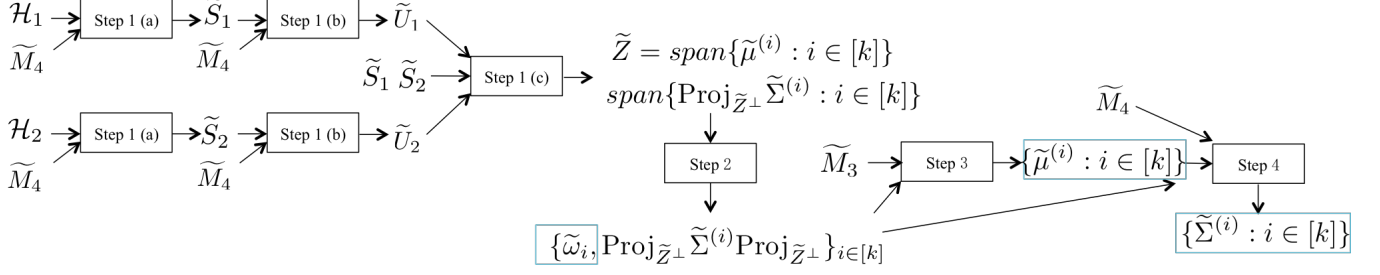
Figure 5: Flow of the algorithm for the general case

G Matrix Perturbation, Concentration Bounds and Auxiliary Lemmas

In this section we collect known results on matrix perturbation and concentration bounds. In general, matrix perturbation bounds are the key for the perturbation lemmas, and concentration bounds are crucial for the smoothed analysis lemmas. We also prove some corollaries of known results that are very useful in our settings.

G.1 Matrix Perturbation Bounds

Given a matrix $\widehat{A} = A + E$ where E is a small perturbation, how does the singular values and singular vectors of A change? This is a well-studied problem and many results can be found in Stewart and Sun Stewart (1977). Here we review some results used in this paper, and prove some corollaries.

Given $\widehat{A} = A + E$, the perturbation in individual singular values can be bounded by Weyl's theorem:

Theorem G.1 (Weyl's theorem)**.** *Given $\widehat{A} = A + E$, we know $\sigma_k(A) - \|E\| \leq \sigma_k(\widehat{A}) \leq \sigma_k(A) + \|E\|$.*

We can also bound the ℓ_2 norm change in singular values by Mirsky's Theorem.

Lemma G.2 (Mirsky's theorem)**.** *Given matrices $A, E \in \mathbb{R}^{m \times n}$ with $m \geq n$, then*

$$\sqrt{\sum_{i=1}^{n}(\sigma_i(A + E) - \sigma_i(A))^2} \leq \|E\|_F.$$

For singular vectors, the perturbation is bounded by Wedin's Theorem:

Lemma G.3 (Wedin's theorem; Theorem 4.1, p.260 in Stewart and Sun (1990))**.** *Given matrices $A, E \in \mathbb{R}^{m \times n}$ with $m \geq n$. Let A have the singular value decomposition*

$$A = [U_1, U_2, U_3] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix} [V_1, V_2]^\top.$$

Let $\widehat{A} = A + E$, with analogous singular value decomposition. Let Φ be the matrix of canonical angles between the column span of U_1 and that of $\widehat{U}_1$, and Θ be the matrix of canonical angles between the column span of V_1 and that of $\widehat{V}_1$. Suppose that there exists a δ such that

$$\min_{i,j} |[\Sigma_1]_{i,i} - [\Sigma_2]_{j,j}| > \delta, \quad and \quad \min_{i,i} |[\Sigma_1]_{i,i}| > \delta,$$

then

$$\| \sin(\Phi) \|^2 + \| \sin(\Theta) \|^2 \leq 2 \frac{\|E\|^2}{\delta^2}.$$

We do not go into the definition of canonical angles here. The only way we will be using this lemma is by combining it with the following:

Lemma G.4 (Theorem 4.5, p.92 in Stewart and Sun (1990))**.** *Let Φ be the matrix of canonical angles between the column span of U and that of $\widehat{U}$, then*

$$\| Proj_{\widehat{U}} - Proj_U \| = \| \sin \Phi \|.$$

As a corollary, we have:

Lemma G.5. *Given matrices $A, E \in \mathbb{R}^{m \times n}$ with $m \geq n$. Suppose that the A has rank k and the smallest singular value is given by $\sigma_k(A)$. Let $\mathcal{S}$ and $\widehat{\mathcal{S}}$ be the subspaces spanned by the first k eigenvectors of A and $\widehat{A} = A + E$, respectively. Then we have:*

$$\| \widehat{S} - \widetilde{S} \| \leq \| Proj_{\widehat{\mathcal{S}}} - Proj_{\mathcal{S}} \| = \| Proj_{\widehat{\mathcal{S}}^\perp} - Proj_{\mathcal{S}^\perp} \| \leq \frac{\sqrt{2}\|E\|_F}{\sigma_k(A)}.$$

Moreover, if $\|E\|_F \leq \sigma_k(A)/\sqrt{2}$ we have $\| \widehat{S} - \widetilde{S} \| \leq \frac{\sqrt{2}\|E\|}{\sigma_k(A)}$.

Proof. We first prove the first inequality:

$$\| \mathrm{Proj}_{\widehat{\mathcal{S}}} - \mathrm{Proj}_{\mathcal{S}} \| = \| 2\widetilde{S}(\widehat{S} - \widetilde{S})^\top + (\widehat{S} - \widetilde{S})(\widehat{S} - \widetilde{S})^\top \| \geq 2\|\widetilde{S}\|\|\widehat{S} - \widetilde{S}\| - \|\widehat{S} - \widetilde{S}\|^2 \geq \|\widetilde{S}\|\|\widehat{S} - \widetilde{S}\| = \|\widehat{S} - \widetilde{S}\|.$$

The equality is because $\mathrm{Proj}_{\mathcal{S}^\perp} = I - \mathrm{Proj}_\mathcal{S}$ so the two differences are the same. The final step follows from Wedin's Theorem and Lemma G.4. $\qquad \square$

Often we need to bound the perturbation of a product of perturbed matrices, where we apply the following lemma:

Lemma G.6. *Consider a product of matrices $A_1 \cdots A_k$, and consider any sub-multiplicative norm on matrix $\| \cdot \|$. Given $\widehat{A}_1, \ldots, \widehat{A}_k$ and assume that $\|\widehat{A}_i - A_i\| \leq \|A_i\|$, then we have:*

$$\| \widehat{A}_1 \cdots \widehat{A}_k - A_1 \cdots A_k \| \leq 2^{k-1} \prod_{i=1}^{k} \|A_i\| \sum_{i=1}^{k} \frac{\|\widehat{A}_i - A_i\|}{\|A_i\|}.$$

The proof of this lemma is straightforward by induction.

Perturbation bound for pseudo-inverse When we have a lowerbound on $\sigma_{min}(A)$, it is easy to get bounds for the perturbation of pseudoinverse.

Theorem G.7 (Theorem 3.4 in Stewart (1977))**.** *Consider the perturbation of a matrix $A \in \mathbb{R}^{m \times n}$: $B = A + E$. Assume that $rank(A) = rank(B) = n$, then*

$$\|B^\dagger - A^\dagger\| \leq \sqrt{2}\|A^\dagger\|\|B^\dagger\|\|E\|.$$

As a corollary, we often use:

Lemma G.8. *Consider the perturbation of a matrix $A \in \mathbb{R}^{m \times n}$: $B = A + E$ where $\|E\| \leq \sigma_{min}(A)/2$. Assume that $rank(A) = rank(B) = n$, then*

$$\|B^\dagger - A^\dagger\| \leq 2\sqrt{2}\|E\|/\sigma_{min}(A)^2.$$

Proof. We first apply Theorem G.7, and then bound $\|A^\dagger\|$ and $\|B^\dagger\|$. By definition we know $\|A^\dagger\| = 1/\sigma_{min}(A)$. By Weyl's theorem $\sigma_{min}(B) \geq \sigma_{min}(A) - \|E\| \geq \sigma_{min}(A)/2$, hence $\|B^\dagger\| = \sigma_{min}(B)^{-1} \leq 2\sigma_{min}(A)^{-1}$. $\qquad \square$

G.2 Lowerbounding the Smallest Singular Value

Gershgorin's Disk Theorem is very useful in bounding the singular values.

Theorem G.9 (Gershgorin's theorem). *Given a symmetric matrix $X \in \mathbb{R}^{k \times k}$, a lower bound on the smallest eigenvalue is given by:*

$$\sigma_{min}(X) \geq \min_{i \in [k]} \left\{ X_{i,i} - \sum_{j \in [k], j \neq i} X_{i,j} \right\}.$$

Sometimes, it is easier to consider the projection of a matrix. Lowerbounding the smallest singular value of a projection will imply the same lowerbound on the original matrix:

Lemma G.10. *Suppose $A \in \mathbb{R}^{m \times n}$, let $P \in \mathbb{R}^{m \times d}$ be a subspace, then $\sigma_k(P^\top A) \leq \sigma_k(A)$.*

Proof. Observe that $(P^\top A)^\top (P^\top A) = A^\top (PP^\top) A \preceq A^\top A$ (because P is a subspace). Therefore the eigenvalues of $(P^\top A)^\top (P^\top A)$ must be dominated by the eigenvalues of $A^\top A$. Then the lemma follows from the definition of singular values. □

As a corollary we have the following lemma:

Lemma G.11. *Let $A \in \mathbb{R}^{m \times n}$ and suppose that $m \geq n$. For any projection $Proj_S$, we have that the singular values are non-increasing after the projection:*

$$\sigma_i(Proj_S(A)) \leq \sigma_i(A), \quad for \ i = 1, \ldots, n.$$

In several places of this work we want to bound the singular value of a matrix, where part of the matrix has a block structure.

Lemma G.12. *For given matrices $B^{(i)} \in \mathbb{R}^{m \times n}$ and $C^{(i)} \in \mathbb{R}^{m \times n'}$ for $i = 1, \ldots, d$. Suppose $md > (n + n'd)$, Define the tall matrix $A \in \mathbb{R}^{md \times (n + dn')}$:*

$$A = \begin{bmatrix} B^{(1)} & C^{(1)} & 0 & \cdots & 0 \\ B^{(2)} & 0 & C^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B^{(d)} & 0 & 0 & \cdots & C^{(d)} \end{bmatrix} = \left[B, \, diag(C^{(i)}) \right].$$

The smallest singular value is bounded by:

$$\sigma_{(n+dn')}(A) \geq \min\{\sigma_n(B), \ \sigma_{n'}(Proj_{(B^{(i)})^\perp} C^{(i)}) : i = 1, \ldots, d\}.$$

Proof. The idea is to break the matrix into two parts $A = Proj_B A + Proj_{B^\perp} A$. Since these two spaces are orthogonal we know $\sigma_{(n+dn')}(A) \geq \min\{\sigma_n(Proj_B A), \sigma_{dn'}(Proj_{B^\perp} A)\}$.

For the first part, clearly $\sigma_n(Proj_B A) \geq \sigma_n(B)$, as B is a submatrix of $Proj_B A$.

For the second part, we actually do the projection to a smaller subspace: for each block we project to the orthogonal subspace of $B^{(i)}$. Under this projection, the block structure is preserved. The dn'-th singular value must be at least the minimum of the n'-th singular value of the blocks. In summary we have:

$$\sigma_{(n+dn')}(A) \geq \min\{\sigma_n(B), \ \sigma_{dn'}(Proj_{B^\perp} diag(C^{(i)}))\}$$
$$\geq \min\{\sigma_n(B), \ \sigma_{dn'}(Proj_{diag((B^{(i)})^\perp)} diag(C^{(i)}))\}$$
$$\geq \min\{\sigma_n(B), \ \sigma_{dn'}(diag(Proj_{(B^{(i)})^\perp} C^{(i)}))\}$$
$$\geq \min\{\sigma_n(B), \ \sigma_{n'}(Proj_{(B^{(i)})^\perp} C^{(i)}) : i = 1, \ldots, d\}.$$

□

Smallest singular value of random matrices In our analysis, we often also want to bound the smallest singular value of a matrix whose entries are Gaussian random variables. Our analysis mostly builds on the following results in random matrix theory.

For a random rectangular matrix, Rudelson and Vershynin (2009) gives the following nice result:

Lemma G.13 (Theorem 1.1 in Rudelson and Vershynin (2009)). *Let $A \in \mathbb{R}^{m \times n}$ and suppose that $m \geq n$. Assume that the entries of A are independent standard Gaussian variable, then for every $\epsilon > 0$, with probability at least $1 - (C\epsilon)^{m-n+1} + e^{-C'n}$, where C, C' are two absolute constants, we have:*

$$\sigma_n(A) \geq \epsilon(\sqrt{m} - \sqrt{n-1}).$$

We will mostly use an immediate corollary of the above lemma with slightly simpler form:

Corollary G.14. *Let $A \in \mathbb{R}^{m \times n}$ and suppose that $m \geq 2n$. Assume that the entries of A are independent standard Gaussian variable, then for every $\epsilon > 0$, and for some absolute constant C, with probability at least $1 - (C\epsilon)^{0.5m}$, we have:*

$$\sigma_n(A) \geq \epsilon\sqrt{m}.$$

This lemma can also be applied to a projection of a Gaussian matrix:

Lemma G.15. *Given a Gaussian random matrix $E \in \mathbb{R}^{m \times n}$, for some set $\mathcal{J} \in [m]$ define $E_J = [E_{[j,:]} : j \in \mathcal{J}]$ and $E_{J^c} = [E_{[j,:]} : j \in [m]/\mathcal{J}]$. Define matrix $S \in \mathbb{R}^{n \times r}$ whose columns are orthonormal. Suppose that the matrix S is an arbitrary function of E_J and is independent of E_{J^c}. Assume that*

$$m - |\mathcal{J}| - r \geq 2n \tag{38}$$

Then for any $\epsilon > 0$, we have that with probability at least $1 - (C\epsilon)^{0.5(m-|\mathcal{J}|-r)}$, for some absolute constant C, the smallest singular value of the projected random matrix is bounded by:

$$\sigma_n(\mathrm{Proj}_{S^\perp}E) \geq \epsilon\sqrt{m - |\mathcal{J}| - r}. \tag{39}$$

Proof. For a matrix $A \in \mathbb{R}^{m \times n}$, define the fixed matrix $P_{J^c} \in \mathbb{R}^{(m-|\mathcal{J}|) \times m}$ such that:

$$\left[[P_{J^c}]_{[:,j]} : j \in \mathcal{J}\right] = 0, \qquad \left[[P_{J^c}]_{[:,j]} : j \in [n]/\mathcal{J}\right] = I_{(m-|\mathcal{J}|) \times (m-|\mathcal{J}|)},$$

which only keeps the coordinates that correspond to $[m]/\mathcal{J}$ of any vector in $\mathbb{R}^m$. Note that

$$\begin{aligned}
\sigma_n(\mathrm{Proj}_{S^\perp}E) &\geq \sigma_n(P_{J^c}(\mathrm{Proj}_{S^\perp}E)) \\
&\geq \sigma_n(\mathrm{Proj}_{(P_{J^c}S)^\perp}P_{J^c}\mathrm{Proj}_{S^\perp}E) \\
&= \sigma_n(\mathrm{Proj}_{(P_{J^c}S)^\perp}P_{J^c}E).
\end{aligned}$$

We justify the last equality below. Note that

$$\mathrm{Proj}_{S^\perp}E = E - \mathrm{Proj}_S E,$$

and note that the columns of $(P_{J^c}\mathrm{Proj}_S E)$ lie in the column span of $P_{J^c}S$, therefore,

$$\begin{aligned}
\mathrm{Proj}_{(P_{J^c}S)^\perp}P_{J^c}\mathrm{Proj}_{S^\perp}E &= \mathrm{Proj}_{(P_{J^c}S)^\perp}P_{J^c}E - \mathrm{Proj}_{(P_{J^c}S)^\perp}(P_{J^c}\mathrm{Proj}_S E) \\
&= \mathrm{Proj}_{(P_{J^c}S)^\perp}P_{J^c}E.
\end{aligned}$$

Finally, note that $P_{J^c}S$, with column rank no more than r, is independent of $P_{J^c}E$, which is a random Gaussian matrix of size $(m-|\mathcal{J}|) \times n$, therefore we have that $\mathrm{Proj}_{(P_{J^c}S)^\perp}P_{J^c}E$ is equivalent to a $(m-|\mathcal{J}|-r) \times n$ random Gaussian matrix. Since (38) is satisfied, we can apply Lemma G.13 and conclude (39) with high probability. $\qquad\square$

However, in the smoothed analysis setting, the matrix we are interested in are often not random Gaussian matrices. Instead they are fixed matrices perturbed by Gaussian variables. We call these "perturbed rectangular matrices", their singular values can be bounded as follows:

Lemma G.16 (Perturbed rectangular matrices)**.** *Let $A \in \mathbb{R}^{m \times n}$ and suppose that $m \geq 3n$. If all the entries of A are independently ρ-perturbed to yield $\widetilde{A}$, then for any $\epsilon > 0$, with probability at least $1 - (C\epsilon)^{0.25m}$, for some absolute constant C, the smallest singular value of $\widetilde{A}$ is bounded below by:*

$$\sigma_n(\widetilde{A}) \geq \epsilon\rho\sqrt{m}.$$

Proof. The idea is to use the previous lemma and project to the orthogonal subspace of A. We have that $\widetilde{A} = A + E$, where $E \in \mathbb{R}^{m \times n}$ is a random Gaussian matrix.

$$\sigma_n(\widetilde{A}) \geq \sigma_n(\text{Proj}_{A^\perp}\widetilde{A}) = \sigma_n(\text{Proj}_{A^\perp}E).$$

Since $m - n > 2n$, we can apply Lemma G.15 to conclude that for any $\epsilon > 0$,

$$\sigma_n(\text{Proj}_{A^\perp}E) \geq \epsilon\rho\sqrt{m},$$

with probability at least $1 - (C\epsilon)^{0.5(m-n)} \leq 1 - (C\epsilon)^{0.25m}$. $\qquad\square$

G.3 Projection of random vectors

In Step 2, we need to bound the norm of a random vector of the form $u \odot v$ after a projection, where u and v are two Gaussian vectors. In order to show this, we apply the result in Vu and Wang (2013) which provides a concentration bound of projection of well-behaved (K-concentrated) random vectors.

First we cite the definition of "K-concentrated" below:

Definition G.17. *A random vector $X = (\xi_1, \xi_2, ..., \xi_n)$ is K-concentrated (where K may depend on n) if there are constants $C, C' > 0$ such that for any convex, 1-Lipschitz function $f : \mathbb{C}^n \to \mathbb{R}$ and for any $t > 0$, we have:*

$$\Pr[|F(X) - med(F(X))| \geq t] \leq C \exp\left(-C'\frac{t^2}{K^2}\right),$$

where $med(\cdot)$ denotes the median of a random variable (choose an arbitrary one if there are many).

Lemma G.18 (Concentration for Random Projections (Lemma 1.2 in Vu and Wang (2013)))**.** *Let v be a K-concentrated random vector in $\mathbb{C}^n$. The entries of v has expected norm 1. Then there are constants $C, C' > 0$ such that the following holds. Let $Proj_S$ be a projection to a d-dimensional subspace in $\mathbb{C}^n$.*

$$\mathbb{P}\left(\left|v^\top Proj_S v - d\right| \geq 2t\sqrt{d} + t^2\right) \leq C\exp(-C'\frac{t^2}{K^2}).$$

In order to apply this lemma in our setting, we need to prove the vectors that we are interested in is K-concentrated:

Lemma G.19. *Conditioned on the high probability event that $\|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}$, the vector $[[E_{[:,i]} \odot E_{[:,j]}]_{s,s'} : s < s']$ is $2\sqrt{n_2}$-concentrated.*

Proof. For any 1-Lipschitz function F on $[[E_{[:,i]} \odot E_{[:,j]}]_{s,s'} : s < s']$, we can define a function $G(E_{[:,i]}, E_{[:,j]}) = F([[E_{[:,i]} \odot E_{[:,j]}]_{s,s'} : s < s'])$ (if $i = j$ then the function G only takes $E_{[:,i]}$ as the variable). Under the assumption that $\|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}$, this new function G is $2\sqrt{n_2}$-Lipschitz.

Now we extend G to G^* when the input $\|E_{[:,i]}\|, \|E_{[:,j]}\| > 2\sqrt{n_2}$. Define the truncation function $\text{trunc}(v) = v$ for $\|v\| \leq 2\sqrt{n_2}$, and $\text{trunc}(v) = 2\sqrt{n_2}v/\|v\|$ for $\|v\| > 2\sqrt{n_2}$. Define the extended function $G^*(E_{[:,i]}, E_{[:,j]}) = G(\text{trunc}(E_{[:,i]}), \text{trunc}(E_{[:,j]}))$, which is still $2\sqrt{n_2}$-Lipschitz since the truncation function is 1-Lipschitz.

Note that for the two Gaussian random vectors $E_{[:,i]}, E_{[:,j]} \sim N(0, I)$, we can apply Gaussian concentration bound in Theorem G.20 on G^*, which implies

$$\mathbb{P}[|G^*(E_{[:,i]}, E_{[:,j]}) - \text{med}(G^*(E_{[:,i]}, E_{[:,j]}))| \geq t] \leq C \exp(-C't^2/4n_2).$$

Since the probability of the event $\|E_{[:,i]}\|, \|E_{[:,j]}\| > 2\sqrt{n_2}$ is very small ($\sim \exp(-\Omega(n_2))$), we have $\delta = |\text{med}(G(E_{[:,i]}, E_{[:,j]})) - \text{med}(G^*(E_{[:,i]}, E_{[:,j]}))|$ in the order of $O(\sqrt{n_2})$. Therefore, for $t \sim \Omega(\sqrt{n_2})$, we have

$$\mathbb{P}[|G^*(E_{[:,i]}, E_{[:,j]}) - \text{med}(G(E_{[:,i]}, E_{[:,j]}))| \geq t] \leq \mathbb{P}[|G^*(E_{[:,i]}, E_{[:,j]}) - \text{med}(G(E_{[:,i]}, E_{[:,j]}))| \geq t - \delta]$$
$$\leq C \exp(-C't^2/4n_2).$$

Finally,

$$\mathbb{P}\left[\left|G(E_{[:,i]}, E_{[:,j]}) - \text{med}(G(E_{[:,i]}, E_{[:,j]}))| \geq t \right| \|E_{[:,i]}\|, \|E_{[:,j]}\| \leq 2\sqrt{n_2}\right]$$
$$\leq \frac{\mathbb{P}[|G^*(E_{[:,i]}, E_{[:,j]}) - \text{med}(G(E_{[:,i]}, E_{[:,j]}))| \geq t]}{\mathbb{P}[\|E_{[:,i]}\| \geq 2\sqrt{n_2} \text{ or } \|E_{[:,i]}\| \geq 2\sqrt{n_2}]}$$
$$\leq C \exp(-C't^2/4n_2).$$

Therefore the random vector $[[E_{[:,i]} \odot E_{[:,j]}]_{s,s'} : s < s']$ is $2\sqrt{n_2}$-concentrated. $\qquad\square$

Theorem G.20 (Gaussian concentration bound). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function which is Lipschitz with constant 1. Consider a random vector $X \sim \mathcal{N}(0, I_n)$. For any $s > 0$ we have*

$$\mathbb{P}\left(\left|f(X) - \mathbb{E}[f(X)]\right| \geq s\right) \leq 2e^{-Cs^2},$$

for all $s > 0$ and some absolute constant $C > 0$.

G.4 Gaussian Chaoses

In Step 2, we want to show that the inner product of two random vectors of the form $< \text{Proj}(u \odot v), \text{Proj}(u \odot v) >$ is small, where u, u' and v, v' are Gaussian vectors. In order to show this, we treat the inner product as a (homogeneous) Gaussian chaos, which is defined to be a homogeneous polynomial over Gaussian random variables[10]. Our analysis builds on the results of many works studying the concentration bound of Gaussian chaoses.

For decoupled Gaussian chaoses, we mostly use the following theorem, which is a simple corollary of Lemma G.22.

---

[10]In fact, the squared norm of projected random vectors considered previously is a special case of Gaussian chaos, and we treat it separately.

Theorem G.21. *Suppose* $a = (a_{i_1,\ldots,i_d})_{1 \le i_1,\ldots,i_d \le n}$ *is a d-indexed array, and* $\|a\|_F$ *denotes its Frobenius norm. Let* $(X_i^{(j)})_{1 \le i \le n, j=1,\ldots,d}$ *be independent copies of* $X \sim \mathcal{N}(0, I_n)$. *For any fixed* $\epsilon > 0$, *with probability at least* $1 - C\exp\left(-C'n^{2\epsilon/d}\right)$,

$$\left| \sum_{i_1,\ldots,i_d=1}^{n} a_{i_1,\ldots,i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)} \right| \le \|a\|_F n^\epsilon.$$

Lemma G.22 (Gaussian chaoses concentration (Corollary 1 in Latała et al. (2006)))**.** *Suppose* $a = (a_{i_1,\ldots,i_d})_{1 \le i_1,\ldots,i_d \le n}$ *is a d-indexed array. Consider a decoupled Gaussian chaos* $G = \sum_{i_1,\ldots,i_d} a_{i_1,\ldots,i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)}$, *where* $X_i^{(k)}$ *are independent copies of the standard normal random variable for all* $i \in [n], k \in [d]$.

$$\mathbb{P}\left(|G| \ge t\right) \le C_d \exp\left(-\frac{1}{C_d} \min_{1 \le k \le d} \min_{(I_1,\ldots,I_k) \in S(k,d)} \left(\frac{t}{\|a\|_{I_1,\ldots,I_k}}\right)^{2/k}\right),$$

where $C_d \in (0, \infty)$ *depends only on d, and* $S(k, d)$ *denotes a set of all partitions of* $\{1, \ldots, d\}$ *into* k *nonempty disjoint sets* $I_1, \ldots, I_k$, *and the norm* $\|\cdot\|_{I_1,\ldots,I_k}$ *is given by:*

$$\|a\|_{I_1,\ldots,I_k} := \sup\left\{ \sum_{i_1,\ldots,i_d} a_{i_1,\ldots,i_d} x_{i_{I_1}}^{(1)} \cdots x_{i_{I_k}}^{(k)} : \sum_{i_{I_1}} (x_{i_{I_1}}^{(1)})^2 \le 1, \ldots, \sum_{i_{I_k}} (x_{i_{I_k}}^{(k)})^2 \le 1 \right\}.$$

Proof. (of Theorem G.21) Apply the inequality:

$$\|a\|_{\{1\},\ldots,\{d\}} \le \|a\|_{I_1,\ldots,I_k} \le \|a\|_{[d]} = \|a\|_F, \quad \forall(I_1,\ldots,I_k) \in S(k,d).$$

For a fixed order d and for any $\epsilon > 0$, apply Lemma G.22 and set $t = n^\epsilon \|a\|_F$. We have that $\mathbb{P}\left(|G| \ge t\right) \le C\exp\left(-C'n^{2\epsilon/d}\right)$, for some constant C, C'. $\qquad \square$

For coupled Gaussian chaoses, namely when $X^{(j)}$'s are identical copies of the same X, we first cite the following decoupling theorem in de la Peña and Montgomery-Smith (1995).

Theorem G.23. *(Decoupling) Let* $(a_{i_1,\ldots,i_d})_{1 \le i_1,\ldots,i_d \le n}$ *be a symmetric d-indexed array such that* $a_{i_1,\ldots,i_d} = 0$ *whenever there exists* $k \ne l$ *such that* $i_k = i_l$. *Let* $X_1, \ldots, X_n$ *be independent random variables and* $(X_i^{(j)})_{1 \le i \le n}$ *for* $j = 1, dots, d$, *be independent copies of the sequence* $(X_i)_{1 \le i \le n}$, *then for all* $t \ge 0$,

$$L_d^{-1} \Pr\left[\left| \sum_{i_1,\ldots,i_d=1}^{n} a_{i_1,\ldots,i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)} \right| \ge L_d t \right] \le \Pr\left[\left| \sum_{i_1,\ldots,i_d=1}^{n} a_{i_1,\ldots,i_d} X_{i_1} \cdots X_{i_d} \right| \ge L_d t \right]$$

$$\le L_d \Pr\left[\left| \sum_{i_1,\ldots,i_d=1}^{n} a_{i_1,\ldots,i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)} \right| \ge L_d^{-1} t \right],$$

where $L_d \in (0, \infty)$ *depends only on d.*

Essentially this theorem shows for a symmetric tensor with no "diagonal" terms, i.e., $a_{i_1,\ldots,i_d} = 0$ whenever there exists $k \ne l$ such that $i_k = i_l$), there is only a constant factor difference between the coupled and decoupled Gaussian chaos distribution.

In most of our applications, we do have symmetric tensors with no "diagonal" terms. However there is one case where we do have diagonal terms, for which we need the following lemma.

Lemma G.24. *Let $(a_{i_1,i_2,i_3})_{1 \leq i_1,\ldots,i_3 \leq n}$ be a symmetric 3-indexed array and let $\|a\|_F$ denote its Frobenius norm. Let $X \sim \mathcal{N}(0, I_n)$, then for any $\epsilon > 0$, with probability at least $1 - Cn\exp(-C'n^{2\epsilon/3})$,*

$$\left| \sum_{i_1,i_2,i_3=1}^{n} a_{i_1,i_2,i_3} X_{i_1} X_{i_2} X_{i_3} \right| \leq 4\|a\|_F n^{0.5+\epsilon}.$$

Proof. The sum of the "diagonal" terms is equal to $3\sum_{i \neq j} a_{i,i,j} X_i^2 X_j + 1/2 \sum_i a_{i,i,i} X_i^3$. Since X_i are independent standard Gaussian random variables, with probability at least $1 - Cn\exp(-C'n^{2\epsilon/3})$ (union bound), $|X_i| \leq n^{\epsilon/3}$ for all $i \in [n]$. Conditioned on this high probability event, the absolute value of the sum is bounded by:

$$
\left| 3\sum_{i \neq j} a_{i,i,j} X_i^2 X_j + \frac{1}{2} \sum_i a_{i,i,i} X_i^3 \right| \leq 3 \sum_{i,j=1}^{n} |a_{i,i,j}||X_j|X_i^2
$$
$$
\leq 3\|(a_{i,i,j})_{1 \leq i,j \leq n}\|_1 n^\epsilon
$$
$$
\leq 3\sqrt{n}\|(a_{i,i,j})_{1 \leq i,j \leq n}\|_F n^\epsilon
$$
$$
\leq 3\|a\|_F n^{0.5+\epsilon}.
$$

By Theorem G.21, we know that with probability at least $1 - C\exp\left(-C'n^{2\epsilon/3}\right)$, the absolute value of the sum of the "non-diagonal" terms is bounded by $\|a\|_F n^\epsilon$. Therefore we can conclude the proof by applying the union bound. $\square$