Theoretical Foundations for Clustering and Screening Heterogeneous and High dimensional Data

by

Yun Wei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied and Interdisciplinary Mathematics)
in the University of Michigan
2020

Doctoral Committee:

Professor Alfred O. Hero, III, Co-Chair,
Associate Professor XuanLong Nguyen, Co-Chair
Profesor Jinho Baik
Professor Bala Rajaratnam, University of California, Davis
Professor Mark Rudelson
Assistant Professor Gongjun Xu

Yun Wei

cloudwei@umich.edu

ORCID iD: 0000-0002-9815-1581

DEDICATION

This manual is dedicated to all my teachers.

ACKNOWLEDGMENTS

I can never overstate my gratitude to my advisors Professor Alfred Hero and Professor XuanLong Nguyen. Without Professor Hero's consistent support since the Fall of 2015, especially the support when I had a health problem and when I was not sure whether to pursue academics, I would not be able to write this thesis. It is also Professor Hero who has led me into research and patiently guide me to gradually become more and more independent in research. Professor Nguyen became my coadvisor in 2018, and his vision and taste in statistics have greatly influenced my latter part of my Ph.D. study. It is Professor Nguyen who led me into a whole new area between geometry and statistics and taught me how to build a career in statistics.

I would like to thank Professor Roman Vershynin for serving as my advisor for a period of my Ph.D. study. I enjoy the wonderful courses high dimensional probability taught by him and benefit from many discussions with him. I would like to thank Professor Bala Rajaratnam for his patience and input during the collaboration and for being on my committee. I would like to thank Professor Mark Rudelson, a member of my committee, for many courses I took from him and some insightful discussions on random geometric graphs, which inspires me to improve results in Section 3.5. I would also like to thank Professor Jinho Baik and Professor Gongjun Xu to be on my committee.

Many thanks go out to my colleagues, both past and present, in the Hero lab and Nguyen group for their support throughout the years. I also want to thank Xinzhou Guo for organizing a reading group on high dimensional probability, which I enjoyed a lot.

Finally, I would like to thank my friends who have helped me throughout my time at the University of Michigan, included but not limited to Zhou Zhou, Qingtang Su, Hao Yuan, Hao Wu, Huajie Qian, Jingsheng Wang, Lu Xia, Feng Wei, Yitong Sun, Ming Zhang, Bobbie Wu, Yuchong Zhang, Jiaqi Li, Weicheng Gu, Yining Lu, Yuanyuan Chen, Zhang Jiang, Yifeng Wang, Rose Chang, Yuliang Xu, Beixi Jia, Xiaofang Jiang, Xueyan Wu. Without you all, the time in Ann Arbor would've been much less enjoyable. Special thanks to many friends who have played the Chinese board game Werewolf with me for making my life colorful and allowing me to rejuvenate myself every week.

TABLE OF CONTENTS

DEDICATION . . . . . **ii**

ACKNOWLEDGMENTS . . . . . **iii**

LIST OF FIGURES . . . . . **vii**

ABSTRACT . . . . . **ix**

Chapter

1 Introduction **1**

1.1 Problem Formulations 2
1.1.1 Mixture of product distribution 2
1.1.2 Screening in high dimensional data 3
1.2 Challenges and Contributions 4
1.2.1 Mixture of product distribution 4
1.2.2 Screening in high dimensional data 5
1.3 Geometric Observations 6
1.4 Outline of the thesis and list of relevant publications/preprints 7

2 Mixture of Product Distribution **9**

2.1 Introduction 9
2.2 Background and overview 16
2.2.1 First-order identifiability and inverse bounds 16
2.2.2 General approach and techniques 18
2.3 Preliminaries 20
2.4 First-order identifiability theory 23
2.4.1 Basic theory 23
2.4.2 Finer characterizations 30
2.5 Inverse bounds for mixtures of product distributions 35
2.5.1 Implications on classical and first-order identifiability 36
2.5.2 Probability kernels in regular exponential family 41
2.5.3 General probability kernels 46
2.5.4 Examples of non-standard probability kernels 48
2.6 Posterior contraction of de Finetti's mixing measures 50
2.6.1 Data are equal-length exchangeable sequences 50
2.6.2 Data are variable-length exchangeable sequences 54
2.7 Sharpness of bounds and minimax theorem 56

2.7.1 Sharpness of inverse bounds . . . 56
2.7.2 Minimax lower bounds . . . 58
2.8 Hierarchical model: kernel P_θ is itself a mixture distribution . . . 60
2.8.1 Bounds on oscillatory integrals . . . 60
2.8.2 Kernel P_θ is a mixture of Gaussian distributions . . . 62
2.8.3 Moment map is injective and its Jacobian is of full column rank . . . 66
2.8.4 Kernel P_θ is mixture of Dirichlet processes . . . 67
2.9 Proofs of lemmas in Section 2.3 . . . 71
2.10 Proofs and auxiliary lemmas of Section 2.4 . . . 73
2.10.1 Additional examples and proofs of results in Section 2.4.1 . . . 73
2.10.2 Proofs in Section 2.4.2 . . . 82
2.10.3 Auxiliary Lemmas . . . 85
2.11 Proofs in Section 2.5 . . . 88
2.11.1 Proofs in Section 2.5.1 . . . 88
2.11.2 Proofs and additional examples in Section 2.5.2 . . . 92
2.12 Proof of inverse bounds for mixtures of product distributions . . . 99
2.12.1 Proof of Theorem 2.5.7 . . . 99
2.12.2 Proof of Theorem 2.5.14 . . . 105
2.13 Proofs and auxiliary lemmas of Section 2.6 . . . 119
2.13.1 Proof of Theorem 2.6.2 . . . 119
2.13.2 Proof of Theorem 2.6.5 . . . 122
2.13.3 Auxiliary Lemmas for Section 2.6 . . . 126
2.14 Proofs in Section 2.7 . . . 130
2.15 Proofs in Section 2.8 . . . 134
2.16 Auxiliary lemmas for Section 2.12.2 . . . 140

3 Screening in High Dimensional Data . . . 145

3.1 Introduction . . . 145
3.2 A unified theorem . . . 147
3.2.1 Framework . . . 147
3.2.2 A unified theorem . . . 149
3.2.3 (τ, κ) sparsity . . . 152
3.2.4 Local normalized determinant . . . 154
3.3 Non-asymptotic compound Poisson approximation . . . 155
3.3.1 Score representations of sample correlation and partial correlation . . . 156
3.3.2 Random pseudo geometric graph . . . 159
3.3.3 Closeness of the distribution of the star subgraph counts to compound Poisson . . . 161
3.3.4 A portmanteau proposition on pairwise total variations . . . 166
3.3.5 Unified convergence: an umbrella theorem . . . 169
3.4 Convergence of moments . . . 172
3.5 Explicit characterizations . . . 175
3.5.1 Explicit characterizations for α_ℓ . . . 175
3.5.2 Explicit characterizations for $\alpha(\ell, r_\rho)$ . . . 178
3.6 Conclusions and discussions . . . 180

3.7 Controlling local normalized determinant by extreme eigenvalues 181
3.8 Proofs in Subsection 3.3.1 and Subsection 3.3.2 182
3.8.1 Proof of Lemma 3.3.2 182
3.8.2 Proof of Lemma 3.3.3 (b) 183
3.9 Proof of Proposition 3.3.6 185
3.9.1 Auxiliary lemmas for Proposition 3.3.6 185
3.9.2 Lemmas on double summations 188
3.9.3 Proof of Proposition 3.3.6 195
3.10 Proofs in Subsection 3.3.4 201
3.10.1 Proof of Lemma 3.3.8 201
3.10.2 Proof of Proposition 3.3.9 (a) 201
3.10.3 Proof of Proposition 3.3.9 (b) 202
3.10.4 Proof of Proposition 3.3.9 (c) 210
3.11 Proofs in Subsection 3.3.5 212
3.11.1 Proof of Lemma 3.3.15 212
3.11.2 Proofs of Lemma 3.3.16 and Lemma 3.3.17 214
3.12 Proofs in Section 3.4 218
3.12.1 Proofs of Lemma 3.4.1 and Proposition 3.4.2 218
3.12.2 Proof of Proposition 3.4.3 223
3.13 Proofs in Section 3.5 230
3.13.1 Proofs of Lemma 3.5.3, Lemma 3.5.4 and Corollary 3.5.5 230
3.13.2 Proof of Lemma 3.5.6 234
3.14 Auxiliary lemmas 238
3.15 Numerical simulations and experiments 245

4 Future Directions 248

BIBLIOGRAPHY 250

LIST OF FIGURES

FIGURE

3.1 A graph with 5 vertices and 5 edges. 149

3.2 Diagram of the correlation graph $\mathcal{G}_0(\boldsymbol{\Sigma})$ for $p = 7$ dimensional distributions with two different 7×7 covariance matrices. The left panel is associated with a block-3 sparse assumption on $\boldsymbol{\Sigma}$. Only the $\tau = 3$ variables in the group inside the left circle are correlated: there is no correlation (edge) between the remaining 4 variables in the right circle and there is no correlation across the two sets of variables in different circles. The right panel is associated with $(\tau, \kappa) = (3, 3)$ sparsity on $\boldsymbol{\Sigma}$, where two additional edges, representing correlations between variables, exist across the two groups. 154

3.3 This graph has the 6 quantities associated to empirical correlation or partial correlation graph as vertices. The 4 solid edges correspond to existence of an direct upper bound of the total variation between two vertices, with the weights respectively correspond to the 4 upper bounds (neglecting constant coefficients) in Proposition 3.3.9. Dash edges correspond to an indirect upper bound of the total variation between vertices, with weights computed from solid path connecting the two vertices. 168

3.4 (a) is a comparison in the log-scale between the upper bound on $d_{TV}(\boldsymbol{\zeta}_{n,2}, \delta_{\{1\}})$ by (3.50) with $\delta = 2$ and the exact value of $d_{TV}(\boldsymbol{\zeta}_{n,2}, \delta_{\{1\}})$ by (3.49). (b) is the plot of the upper bound on $d_{TV}(\boldsymbol{\zeta}_{n,2}, \delta_{\{1\}})$ as a function of n for different values of δ from 2 to 7. 178

3.5 The solid circle represents the unit Euclidean ball B_2^n in $\mathbb{R}^n$ while the dash circle represents the unit ball centered at $\tilde{\boldsymbol{u}}_1$. Their intersection is the green region, which is contained in contained in the ball with center at $\tilde{\boldsymbol{u}}_1/2$ and with radius $\sqrt{1 - \left(\frac{\|\tilde{\boldsymbol{u}}_1\|_2}{2}\right)^2}$. 231

3.6 $\mathbf{0}$ is the origin in $\mathbb{R}^{n-2}$ and $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_4, \boldsymbol{z}_5$ are on S^{n-2}. $\boldsymbol{z}_3$ is the midpoint of z_1 and z_2, while z_4 is the midpoint of the shortest arc on S^{n-2} connecting $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. $\boldsymbol{z}_5$ is one of the two intersection points of the boundary $\mathrm{SC}(r, \boldsymbol{z}_1)$ and the boundary of $\mathrm{SC}(r, \boldsymbol{z}_2)$. The angle between line segment $\mathbf{0}\boldsymbol{z}_4$ and $\mathbf{0}\boldsymbol{z}_5$ is θ. 238

3.7 $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are the normalized vector of $\bar{\boldsymbol{z}}_1$ and $\bar{\boldsymbol{z}}_2$ respectively. 243

3.8 The vertical axis of (a) is $d_{\mathrm{TV}}(N_{V_1}^{(k)}, \mathrm{CP}(\lambda_{20,1}(1), \boldsymbol{\zeta}_{20,1}))$ as in Theorem 3.2.4 and that of (b) is $d_{\mathrm{TV}}(N_{V_1}^{(k)}, \mathrm{CP}(\lambda_{p,20,1,\rho}, \boldsymbol{\zeta}_{20,1,\rho}))$ as in Theorem 3.3.11. For both plots the samples are generated according to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being a $(\tau = p^{0.6}, \kappa = p^{0.8})$ sparse matrix for each p. The parameters are $n = 20$, $\delta = 1$ and the threshold ρ is chosen according to (3.12) with $e_{n,\delta} = 1$. The blue curve is for the empirical correlation graph (k=$\boldsymbol{R}$) and the red curve is for the empirical partial correlation graph (k=$\boldsymbol{P}$). Note since $\delta = 1$, $\boldsymbol{\zeta}_{20,1} = \delta_{\{2\}} = \boldsymbol{\zeta}_{20,1,\rho}$, by Example 3.5.1. As demonstrated by the plots, for both empirical correlation and partial correlation graphs, the total variations in (a) decrease very slowly while the total variations in (b) converge to 0 very fast, which has been analytically discussed in Remark 3.3.14. 246

3.9 The vertical axis of (a) is $d_{\mathrm{TV}}(N_{V_\delta}^{(k)}, \mathrm{Pois}(\frac{(e_{n,\delta})^\delta}{\delta!}))$, where we replaced $\mathrm{CP}(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta})$ in Theorem 3.2.4 by its approximation $\mathrm{Pois}(\frac{(e_{n,\delta})^\delta}{\delta!})$ as discussed in Subsection 3.5.1. The vertical axis of (b) is $d_{\mathrm{TV}}(N_{V_\delta}^{(k)}, \mathrm{Pois}(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta))$, where we replaced $\mathrm{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ in Theorem 3.3.11 by its approximation $\mathrm{Pois}(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta)$ as discussed in Subsection 3.5.2. For both plots the samples are generated according to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being a $(\tau = p^{0.6}, \kappa = p^{0.8})$ sparse matrix for each p. The parameters are $n = 35$, $\delta = 2$ and the threshold ρ is chosen according to (3.12) with $e_{n,\delta} = 1$. Note the distributions of the increment $\boldsymbol{\zeta}_{35,2}$ in (a) and $\boldsymbol{\zeta}_{35,2,\rho}$ in (b) are both replaced by $\delta_{\{1\}}$ since $n = 35$ is sufficiently large for $\delta = 2$ as indicated by Figure 3.4 (b). That is, the number of samples $n = 35$ is large enough for Corollary 3.5.5 (a) and Lemma 3.5.6 (c) to be effective. The blue curve is for the empirical correlation graph (k=$\boldsymbol{R}$) and the red curve is for the empirical partial correlation graph (k=$\boldsymbol{P}$). As demonstrated by the plots, for both empirical correlation and partial correlation graphs, the total variations in (a) decrease very slowly while the total variations in (b) converge to 0 very fast. The fast convergence in Figure 3.9 (b) verifies the validity of using Poisson distribution $\mathrm{Pois}(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta)$ to approximate the distribution of random quantities in $\{N_i^{(k)} : k = \boldsymbol{R}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ for large n. The extremely slow decrease in Figure 3.9 (a) is due to the slow convergence of Theorem 3.2.4, which has been extensively discussed in Remark 3.3.14. This specific example indicates the slow convergence of Theorem 3.2.4 is due to slow convergence of $\lambda_{p,n,\delta,\rho} \to \lambda_{n,\delta}$ since the distribution of increments in this large n case are both close to $\delta_{\{1\}}$. 247

ABSTRACT

This thesis is devoted to the study of two problems in statistics which involve complex data structure of high heterogeneity or large scale. Heterogeneous data arises when each data sample can come from a multiplicity of distributions, creating a population distribution that is a finite mixture of these distributions. The large scale may be due to large data samples, or the large number of variables related to the samples, or the imagined infinite dimensions in nonparametric or functional data. In this thesis, we establish theoretical results in both problems, based on techniques from harmonic analysis, random matrix theory, random geometric graphs, and Stein's method.

Mixtures of product distributions are a powerful device for learning about heterogeneity within data populations. In this class of latent structure models, de Finetti's mixing measure plays the central role for describing the uncertainty about the latent parameters representing heterogeneity. In the first part of this thesis posterior contraction theorems for de Finetti's mixing measure arising from finite mixtures of product distributions will be established, under the setting the number of exchangeable sequences of observed variables increases while sequence length(s) may be either fixed or varied. The role of both the number of sequences and the sequence lengths will be carefully examined. In order to obtain concrete rates of convergence, a first-order identifiability theory for finite mixture models and a family of sharp inverse bounds for mixtures of product distributions will be developed via a harmonic analysis of such latent structure models. This theory is applicable to broad classes of probability kernels composing the mixture model of product distributions for both continuous and discrete domain $\mathfrak{X}$. Examples of interest include the case the probability kernel is only weakly identifiable in the sense of [HN16a], the case where the kernel is itself a mixture distribution as in hierarchical models, and the case the kernel may not have a density with respect to a dominating measure on an abstract domain $\mathfrak{X}$ such as Dirichlet processes.

An important problem in large scale inference is the identification of variables that have large correlations or partial correlations with at least one other variable. Recent work in correlation screening has yielded breakthroughs in the ultra-high dimensional setting when the sample size n is fixed and the dimension $p \to \infty$ (see [HR12]). Despite these advances, the correlation screening framework suffers from some serious practical, methodological and theoretical deficiencies. For instance, theoretical safeguards for partial correlation screening requires that the population covariance matrix be block diagonal. This block sparsity assumption is however highly restrictive in

numerous practical applications. As a second example, results for correlation and partial correlation screening framework requires the estimation of dependence measures or functionals, which can be highly prohibitive computationally, rendering the framework impractical and unappealing in the very setting it is designed for. In the second part of this thesis, we propose a unifying approach to correlation and partial correlation screening which specifically goes beyond the block diagonal correlation structure, thus yielding a methodology that is suitable for modern applications. By making insightful connections to random geometric graphs, total number of highly correlated or partial correlated variables are shown to have a novel compound Poisson limit, and are obtained for both the finite p case and when $p \to \infty$. Our approach also obviates the need to estimate dependence measures rendering the framework readily scalable. The unifying framework also demonstrates an important duality between correlation and partial correlation screening with important theoretical and practical consequences.

CHAPTER 1

Introduction

Twenty-first-century technology and science confront applied mathematicians and statisticians with complex data that requires geometric perspectives. Heterogeneous data arises when each data sample can come from a multiplicity of distributions, creating a population distribution that is a finite mixture of these distributions. The large scale may be due to large data samples, or the large number of variables related to the samples, or the imagined infinite dimensions in nonparametric or functional data.

In broad terms, this thesis lies at the intersection of Probability, Geometry and Statistics. More specifically, this thesis addresses these issues of complex heterogeneity and large scale from probabilistic and geometric perspectives. From a methodological standpoint, this thesis analyze models based on nonparametric and graphical statistics, with a particular concern for the issues that arise in exchangeable data or high dimensional data.

The two topics that are the focus of this thesis are as follows. The first topic is to address heterogeneous data by the mixture of product distribution (exchangeable data), i.e. mixture model where each component consists of samples from repeated measurements. The second topic addresses variable screening via thresholding when the variables are of high dimension. These two problems arise in many practical applications, like Latent Dirichlet Allocation in topic modeling, and covariance selection in Gaussian graphical models. The problems are challenging since the data structures are complicated and beyond classical statistical settings: they either involve exchangeable structures instead of i.i.d. structures or involve extreme high dimensional structures while the number of samples is finite and fixed. In this thesis, we develop fundamental theoretical results on parameter estimation and hypothesis testing that requires characterizing the distributions of statistical quantities of interest. The thesis utilizes diverse proof techniques in harmonic analysis, random matrix theory, random geometric graphs, and Stein's method. The common feature of this thesis is the geometric perspective that guides the analysis and provides intuition for interpreting our results.

1.1 Problem Formulations

In this section we present concise (and simplified) problem formulations for the two problems studied in this thesis, mixture of product distribution and screening in high dimensional data.

1.1.1 Mixture of product distribution

Consider a family of probability distributions $\{P_\theta\}_{\theta\in\Theta}$ on measurable space $(\mathfrak{X}, \mathcal{A})$, where θ is the parameter of the family and $\Theta \subset \mathbb{R}^q$ is the parameter space. For $N \in \mathbb{N}$, the N-product probability family is denoted by $\{P_{\theta,N} := \bigotimes^N P_\theta\}_{\theta\in\Theta}$ on $(\mathfrak{X}^N, \mathcal{A}^N)$, where $\mathcal{A}^N$ is the product sigma algebra. Given a mixing measure $G = \sum_{i=1}^k p_i\delta_{\theta_i} \in \mathcal{E}_k(\Theta)$, the mixture of N-product distributions induced by G is given by

$$P_{G,N} = \sum_{i=1}^k p_i P_{\theta_i,N}.$$

Each exchangeable sequence $X^i_{[N]} = (X_{i1}, \ldots, X_{iN})$, for $i = 1, \ldots, m$, is an independent sample distributed according to $P_{G,N}$. It is easy to see that the sequence $X^i_{[N]}$ is exchangeable for each fixed i. Due to the role they play in the composition of distribution $P_{G,N}$, we also refer to $\{P_\theta\}_{\theta\in\Theta}$ as a family of *probability kernels* on $(\mathfrak{X}, \mathcal{A})$. The probability kernel and the number of components k is known, and the goal is to estimate the mixing proportions p_i, and component parameters θ_i. The parameters of interest are always encapsulated by a discrete mixing measure $G \in \mathcal{E}_k(\Theta)$, the space of discrete measures with k distinct support atoms residing in a set $\Theta \subset \mathbb{R}^q$.

Given m independent sequences of exchangeable observations of equal length N, $X^i_{[N]} = (X_{i1}, \cdots, X_{iN}) \in \mathfrak{X}^N$ for $i = 1, 2, \cdots, m$. Each sequence $X^i_{[N]}$ is assumed to be a sample drawn from a mixture of $N-$product distributions $P_{G,N}$ for some "true" mixing measure $G = G_0 \in \mathcal{E}_{k_0}(\Theta)$. A Bayesian statistician endows upon $(\mathcal{E}_{k_0}(\Theta), \mathcal{B}(\mathcal{E}_{k_0}(\Theta)))$ a prior distribution Π and obtains the posterior distribution $\Pi(dG|X^1_{[N]}, \ldots, X^m_{[N]})$ by Bayes' rule, where $\mathcal{B}(\mathcal{E}_{k_0}(\Theta))$ is the Borel sigma algebra w.r.t. W_1 distance, the L^1 Wassertein distance. We study the asymptotic behavior of this posterior distribution as the amount of data $m \times N$ tend to infinity.

It is also customary to express the above Bayesian model in the following hierarchical fashion:

$$\begin{aligned} &G \sim \Pi, \quad \theta_1, \theta_2, \cdots, \theta_m | G \overset{i.i.d.}{\sim} G \\ &X_{i1}, X_{i2}, \cdots, X_{iN} | \theta_i \overset{i.i.d.}{\sim} P_{\theta_i} \quad \text{for } i = 1, \cdots, m. \end{aligned}$$

As above, the m data sequences are denoted by $X^i_{[N]} = (X_{i1}, X_{i2}, \cdots, X_{iN}) \in \mathfrak{X}^N$ for $i = 1, 2, \cdots, m$.

We will show in Chapter 2 that the posterior distribution $\Pi(\cdot|X^1_{[N]}, X^2_{[N]}, \ldots, X^m_{[N]})$ contracts around the truth G_0, of which the precise meaning will be made clear later, if $X^i_{[N]} \overset{i.i.d.}{\sim} p_{G_0,N}$ and $m \to \infty$.

1.1.2 Screening in high dimensional data

The objective is to reliably extract summary statistics on topological properties of a dependency graph based on a sample correlation or inverse correlation matrix. Such properties include edges and vertex degree, among others.

Available is a matrix of multivariate samples

$$\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(n)}]^T = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}, \tag{1.1}$$

where $\{\boldsymbol{x}^{(i)}\}_{i=1}^n \subset \mathbb{R}^p$ are samples from a p-dimensional distribution. The setting is the ultra high dimensional regime where n is far less than p and the regime that n and p are both finite.

The sample mean is

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^{(i)} = \frac{1}{n}\boldsymbol{X}^T\mathbf{1}$$

and the sample covariance matrix $\boldsymbol{S}$ is

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}})(\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}})^T = \frac{1}{n-1}(\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^T)^T(\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^T). \tag{1.2}$$

The sample correlation matrix $\boldsymbol{R}$ is defined as:

$$\boldsymbol{R} = \mathbf{diag}(\boldsymbol{S})^{-\frac{1}{2}}\boldsymbol{S}\mathbf{diag}(\boldsymbol{S})^{-\frac{1}{2}}, \tag{1.3}$$

where $\mathbf{diag}(\boldsymbol{A})$ for a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the diagonal part of $\boldsymbol{A}$ and $\boldsymbol{B}^{-1/2}$ for a diagonal matrix $\boldsymbol{B}$ is a diagonal matrix by raise every diagonal element of $\boldsymbol{B}$ to the power $-1/2$. Since $\boldsymbol{R}$ might not be invertible, we define $\boldsymbol{R}^\dagger$ as the Moore-Penrose pseudo-inverse of $\boldsymbol{R}$ and define the sample partial correlation matrix $\boldsymbol{P}$ by

$$\boldsymbol{P} = \mathbf{diag}(\boldsymbol{R}^\dagger)^{-\frac{1}{2}}\boldsymbol{R}^\dagger\mathbf{diag}(\boldsymbol{R}^\dagger)^{-\frac{1}{2}}. \tag{1.4}$$

The (partial) correlated graph is obtained by thresholding ($\boldsymbol{P}$) $\boldsymbol{R}$ at a certain threshold ρ to obtain an adjacency matrix of the graph. For $\delta \geq 2$, let $N^{(\boldsymbol{R})}_{V_\delta}$ ($N^{(\boldsymbol{P})}_{V_\delta}$) be the number of variables that have sample correlation (partial correlation) above threshold ρ or less than $-\rho$ with at least δ other variables. These variables are connected to at least δ other variables in the graph. For $\delta = 1$, define

$N_{V_\delta}^{(\boldsymbol{R})}$ ($N_{V_\delta}^{(\boldsymbol{P})}$) to be the twice number of edges in the corresponding graph. The goal is to study the distribution of the random quantities $N_{V_\delta}^{(\boldsymbol{R})}$ ($N_{V_\delta}^{(\boldsymbol{P})}$) in the regime n, p are both finite or the regime $p \to \infty$ while n remains fixed.

We will show in Chapter 3 that their distributions are all approximately a compound Poisson, where the parameters of the compound Poisson characterized in terms of the parameters of associated random geometric graphs.

1.2 Challenges and Contributions

In this section we provide the background and challenges for the problems considered in this thesis. The contribution of this thesis on those two problems is summarized at the end of each subsection.

1.2.1 Mixture of product distribution

Latent structure models with many observed variables are among the most powerful and widely used tools in statistics for learning about heterogeneity within data population(s). An important canonical example of such models is the mixture of product distributions, which may be motivated by de Finetti's celebrated theorem for exchangeable sequences of random variables [Ald85, Kal06]. Before the efficiency question can be addressed, one must consider the issue of (classical) identifiability: under what conditions does the data distribution $P_{G,N}$ uniquely identify G? I.e., when is the map $G \mapsto P_{G,N}$ injective? This question has been of great interest to a number of authors [Tei67, EHN05, HNPE05], with decisive results obtained recently by [AMR09]. Their results are quite general, and apply to the case where the observed variables $X_1, \ldots, X_N$ are conditionally independent but not necessarily identically distributed given θ. Here, the condition is in the form of $N \geq n_0$, for some natural constant $n_0 \geq 1$. We shall refer to n_0 as (minimal) *zero-order identifiable length*, or 0-identifiable length for short (a formal definition will be given later). However, their results do not apply in our setup where the data are conditionally independent and identically distributed.

Partial answers to estimating the mixing measure G were obtained in several settings of mixtures of product distributions. [HT00] proposed to discretize data so that the model in consideration becomes a finite mixture of product of identical binomial or multinomial distributions, but they only consider estimate the mixing propositions p_i. Restricting to this class of models, a maximum likelihood estimator was applied, and a standard asymptotic analysis establishes root-m rate for mixing proportion estimates. [HZ03, HNPE05] investigated a number of nonparametric estimators for G, and obtained the root-m convergence rate for both mixing proportion and component parameters in the setting of $k = 2$ mixture components under suitable identifiability conditions.

It seems challenging to extend their method and theory to a more general model setting, e.g., $k > 2$. Moreover, while the role of N on identifiability was discussed, no result on the effect of N on parameter estimation efficiency seems to be available. Recently, [Ngu16, Ngu15] studied the posterior contraction behavior of several classes of Bayesian hierarchical model (including hierarchical Dirichlet processes) under an analogous setting where the sample size is specified by m sequences of N observations. His approach requires that both m and N tend to infinity and thus cannot be applied to our present setting where N may be fixed.

In Chapter 2 we shall present a parameter estimation theory for general classes of finite mixtures of product distributions. An application of this theory will be posterior contraction theorems established for a standard Bayesian estimation procedure, according to which the de Finetti's mixing measure G tends toward the truth G_0, as m tends to infinity, under suitable conditions. We established that as soon as N is sufficiently large, under the Bayesian estimation procedure, the supporting atoms θ_i of G converge toward their true values at the rate bounded from above by $(\ln(mN)/(mN))^{1/2}$. Meanwhile, the mixing probabilities p_j converge toward their true values at the rate bounded by $(\ln(mN)/m)^{1/2}$. Note mN is the total volume of data. In plain terms, we may say that with finite mixtures of product distributions, the posterior inference of atoms of each individual mixture component receives the full benefit of "borrowing strength" across sampled sequences; while the mixing probabilities gain efficiency from only the number of such sequences. This appears to be the first work in which such a posterior contraction theorem is established for de Finetti's mixing measure arising from finite mixtures of product distributions.

1.2.2 Screening in high dimensional data

In Chapter 3 we consider the problem of screening n independent and identically distributed p-variate samples for variables that have high correlation or high partial correlation with at least one other variable in the ultra-high dimensional regime when the sample size $n \leq C_0 \ln p$.[1] In the screening framework one applies a threshold to the sample correlation matrix or the sample partial correlation matrix to detect variables with at least one significant correlation, with the threshold aiming to separate signal from noise. Correlation and partial correlation screening in ultra-high dimensions have become increasingly important in many modern applications as the per-sample cost of collecting high dimensional data is much more costly than per-variable cost. For example, in biomedical settings the cost of high throughput technology, like oligonucleotide gene microchips and RNAseq assays is decreasing, while the cost of biological samples is not decreasing at the same rate [HR15b]. In such situations p is much larger than n.

The ultra-high dimensional regime when $n \leq C_0 \ln p$ is very challenging since the number of

[1] Here C_0 is some universal constant satisfying $C_0 \geq 1$. A "universal constant" or "absolute constant", is a constant that does not depend on any model parameter.

samples is insufficient to apply many (if not most) reliable statistical methods. For example, one way to undertake partial correlation screening is to first estimate the population covariance matrix, then obtain the inverse, from which a partial correlation matrix can be estimated. However, to get a reliable estimate of a general covariance matrix, the number of samples n must be at least $O(p)$ as shown in Section 5.4.3. in [Ver12]. Even if the covariance matrix has a special structure like sparsity, covariance estimation requires a number of samples of order $O(\ln p)$ [RBLZ08].

While estimating the covariance matrix or partial correlation matrix is challenging in ultra-high dimensions, recent work has shown that it is possible to accurately test the number of highly (partial) correlated variables under a false positive probability; in particular the probability that a variable is highly (partially) correlated with at least one other variable [HR11, HR12]. In this thesis we show that the sparsity assumptions in [HR11, HR12] are overly restrictive and can be relaxed. We also correct an error in the proof of one of the theorems in [HR12].

In Chapter 3 we propose a novel unifying framework for correlation and partial correlation screening that delivers a practical and scalable methodology in the ultra-high dimensional regime, which is simultaneously armed with theoretical safeguards. By making novel and insightful connections to random pseudo geometric graphs we demonstrate that the distribution of the number of discoveries tends to a compound Poisson limit. Specifically, let $\boldsymbol{R}$, $\boldsymbol{P}$ denote respectively the sample correlation matrix and sample partial correlation matrix. $N_{V_\delta}^{(\boldsymbol{R})}$ ($N_{V_\delta}^{(\boldsymbol{P})}$) be the number of variables that have sample correlation (partial correlation) above threshold ρ or less than $-\rho$ with at least δ other variables. We show that, under some sparse assumption on the covariance matrix, as long as the threshold ρ is chosen to satisfy that $(1-\rho)^{(n-2)/2}p^{1+1/\delta}$ converges to some constant,

$$N_{V_\delta}^{(k)} \rightarrow \text{ compound Poisson} \qquad \text{in distribution}$$

as $p \rightarrow \infty$ for fixed n and δ, where $k \in \{\boldsymbol{R}, \boldsymbol{P}\} \in \mathbb{R}^{p\times p}$. The parameters of the compound Poisson are also characterized and they are in terms of random geometric graphs. We further established that when n or δ are suitably large, the limit compound Poisson is approximately a Poisson. To the best of our knowledge, such a novel limit has not previously appeared in the correlation screening setting.

1.3 Geometric Observations

In this section some key geometric observations that underpin the results in this thesis are discussed. All these observations involve geometric perspectives and are not only conceptually important but also play important roles in the proofs.

Mixture of product distribution (Chapter 2) The parameters of interest are $\{\{(p_i, \theta_i)\}_{i=1}^k | \sum_i p_i = 1, \theta_i \in \Theta\}$. There are many different ways to represent such parameters, and the advantage to

represent it as a discrete measure $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ as in Section 1.1.1 is to use the Wasserstein distance, which is well studied and convergence in Wasserstein distance automatically implies convergence in p_i and θ_i. The key to proving the posterior contraction rate is to study the geometric property of the map from the space of mixing measure to space of mixture density. More specifically, by showing the map is coercive the convergence of mixture density implies convergence of mixing measures. That is, the heart of the matter lies in the establishment of a collection of general *inverse bounds*, the type of inequalities of the form

$$D_N(G, G_0) \leq C(G_0) V(P_{G,N}, P_{G_0,N}),$$

where $D_N(G, G_0)$ is a variant distance of Wasserstein distance to be discussed in detail in Section 2.3.

Note that (2.2) provides an upper bound on distance D_N of mixing measures in terms of the variational distance between the corresponding mixture of N-product distributions. Inequalities of this type allow one to transfer the convergence (and learning rates) of a data population's distribution into that of the corresponding distribution's parameters (therefore the term "inverse bounds").

Screening in high dimensional data (Chapter 3) One may decompose the sample correlation matrix $\boldsymbol{R} = \boldsymbol{U}^T \boldsymbol{U}$, where $\boldsymbol{U} \in \mathbb{R}^{n-1 \times p}$ satisfies: $\boldsymbol{u}_i$, the i-th column of $\boldsymbol{U}$, has unit Euclidean norm for each $1 \leq i \leq p$. Then the event that the absolute value of the sample correlation between the i-th variable and the j-th variable is above the threshold ρ, is

$$\{|R_{ij}| \geq \rho\} = \{\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 \leq \sqrt{2(1-\rho)}\} \cup \{\|\boldsymbol{u}_i + \boldsymbol{u}_j\|_2 \leq \sqrt{2(1-\rho)}\}.$$

If one identifies there is an edge between the i-th variable (or $\boldsymbol{u}_i$) and the j-th variable (or $\boldsymbol{u}_j$) when the above event holds, then the random graph with vertexes $\{\boldsymbol{u}_i\}_{i=1}^{p}$ are similar to a random Euclidean geometric graph [Pen03]. A key difference, however, is that the vertices of this random graph are not independent and they lie on the unit sphere instead of the whole Euclidean space, a case not covered by the classical Euclidean geometric graph theory of [Pen03]. This observation motives the conclusion and the proof for results on $N_{V_\delta}^{\boldsymbol{R}}$, the number of highly correlated variables.

1.4 Outline of the thesis and list of relevant publications/preprints

In Chapter 2, parameters estimation in the mixture of product distribution is discussed and is based on the following paper.

- Yun Wei, XuanLong Nguyen, Convergence of de Finetti's mixing measure in latent structure models for observed exchangeable sequences, *Annals of Statistics* (under review), 103 pages,

2020

In Chapter 3, variable screening in high dimensional data is discussed and is based on the following paper.

- Yun Wei, Alfred Hero and Bala Rajaratnam, Correlation and partial correlation screening in dimension with relaxed sparsity conditions, to be submitted to *Annals of Statistics*, 70+ pages

Chapter 2 and Chapter 3 are both self-contained, and the readers can read the two chapters in arbitrary order.

CHAPTER 2

Mixture of Product Distribution

2.1 Introduction

Latent structure models with many observed variables are among the most powerful and widely used tools in statistics for learning about heterogeneity within data population(s). An important canonical example of such models is the mixture of product distributions, which may be motivated by de Finetti's celebrated theorem for exchangeable sequences of random variables [Ald85, Kal06]. The theorem of de Finetti states roughly that if $X_1, X_2, \ldots$ is an infinite exchangeable sequence of random variables defined in a measure space $(\mathfrak{X}, \mathcal{A})$, then there exists a random variable θ in some space Θ, where θ is distributed according to a probability measure G, such that $X_1, X_2, \ldots$ are conditionally i.i.d. given θ. Denote by P_θ the conditional distribution of X_i given θ, we may express the joint distribution of a N-sequence $X_{[N]} := (X_1, \ldots, X_N)$, for any $N \geq 1$, as a mixture of product distributions in the following sense: for any $A_1, \ldots, A_N \subset \mathcal{A}$,

$$P(X_1 \in A_1, \ldots, X_N \in A_N) = \int \prod_{n=1}^{N} P_\theta(X_n \in A_n) G(d\theta).$$

The probability measure G is also known as de Finetti's mixing measure for the exchangeable sequence. It captures the uncertainty about the latent variable θ, which describes the mechanism according to which the sequence $(X_i)_i$ is generated via P_θ. In other words, the de Finetti's mixing measure G can be seen as representing the heterogeneity within the data populations observed via sequences $X_{[N]}$. A statistician typically makes some assumption about the family $\{P_\theta\}_{\theta \in \Theta}$, and proceeds to draw inference about the nature of heterogeneity represented by G based on data samples $X_{[N]}$.

In order to obtain an estimate of mixing measure G, one needs multiple copies of the exchangeable sequences $X_{[N]}$. As mentioned, some assumption will be required of the probability distributions P_θ, as well as the mixing measure G. Throughout this chapter it is assumed that the map $\theta \mapsto P_\theta$ is injective. Moreover, we will confine ourselves to the setting of exact-fitted finite mixtures, i.e., G is assumed to be an element of $\mathcal{E}_k(\Theta)$, the space of discrete measures with k distinct

supporting atoms on Θ, where Θ is a subset of $\mathbb{R}^q$. Accordingly, we may express $G = \sum_{j=1}^k p_j \delta_{\theta_j}$. We may write the distribution for $X_{[N]}$ in the following form, where we include the subscripts G and N to signify their roles:

$$P_{G,N}(X_1 \in A_1, \ldots, X_N \in A_N) = \sum_{j=1}^{k} p_j \Big\{ \prod_{n=1}^{N} P_{\theta_j}(X_n \in A_n) \Big\}. \tag{2.1}$$

Note that when $N = 1$, we are reduced to the standard formulation of a mixture distribution $P_G := P_{G,1} = \sum_{j=1}^k p_j P_{\theta_j}$. Due to the role they play in the composition of distribution $P_{G,N}$, we also refer to $\{P_\theta\}_{\theta \in \Theta}$ as a family of *probability kernels* on $\mathfrak{X}$. Given m *independent* copies of exchange sequences $\{X^i_{[N_i]}\}_{i=1}^m$ each of which is respectively distributed according to P_{G,N_i} given in (2.1), where N_i denotes the possibly variable length of the i-th sequence. The primary question of interest in this chapter is the efficiency of the estimation of the true mixing measure $G = G_0 \in \mathcal{E}_k(\Theta)$, for some known $k = k_0$, as sample size $(m, N_1, \ldots, N_m)$ increases in a certain sense.

Before the efficiency question can be addressed, one must consider the issue of (classical) identifiability: under what conditions does the data distribution $P_{G,N}$ uniquely identify G? I.e., when is the map $G \mapsto P_{G,N}$ injective? This question has occupied the interest of a number of authors [Tei67, EHN05, HNPE05], with decisive results obtained recently by [AMR09]. Their results are quite general, and apply to the case where the observed variables $X_1, \ldots, X_N$ are conditionally independent but not necessarily identically distributed given θ. Here, the condition is in the form of $N \geq n_0$, for some natural constant $n_0 \geq 1$. We shall refer to n_0 as (minimal) *zero-order identifiable length*, or 0-identifiable length for short (a formal definition will be given later). In particular, when $\mathfrak{X} = \mathbb{R}^d$, for some $d \geq 1$, suppose the family of probability measures $\{P_\theta\}_{\theta \in \Theta}$ are linearly independent and absolutely continuous with respect to the Lebesgue measure on $\mathfrak{X}$, then $P_{G,N}$ uniquely identifies G as soon as $N \geq 3$. (Note that in the conditional i.i.d. setting, the linear independence condition immediately entails that identifiability holds for all $N \geq 1$. However, the situation is unclear when the linear independence condition is not satisfied). When domain $\mathfrak{X}$ is a finite set, then identifiability is achieved up to a Lebesgue measure-zero set of $\theta \in \Theta$, provided that $N \geq \lceil 2 \log_{|\mathfrak{X}|} k_0 \rceil + 1$.

Drawing from existing identifiability results, it is quite apparent that the observed sequence length N (or more precisely, $N_1, \ldots, N_m$, in case of variable length sequences) must play a crucial role in the estimation of mixing measure G, in addition to the number m of sequences. Moreover, it is also quite clear that in order to have a consistent estimate of $G = G_0$, the number of sequences m must tend to infinity, whereas N may be allowed to be fixed. It remains an open question as to the precise roles m and N play in estimating G and on the different types of mixing parameters: the

component parameters (atoms θ_j) and mixing proportions (probability mass p_j), and the rates of convergence of a given estimation procedure.

Partial answers to this question were obtained in several settings of mixtures of product distributions. [HT00] proposed to discretize data so that the model in consideration becomes a finite mixture of product of identical binomial or multinomial distributions. Restricting to this class of models, a maximum likelihood estimator was applied, and a standard asymptotic analysis establishes root-m rate for mixing proportion estimates. [HZ03, HNPE05] investigated a number of nonparametric estimators for G, and obtained the root-m convergence rate for both mixing proportion and component parameters in the setting of $k = 2$ mixture components under suitable identifiability conditions. It seems challenging to extend their method and theory to a more general model setting, e.g., $k > 2$. Moreover, while the role of N on identifiability was discussed, no result on the effect of N on parameter estimation efficiency seems to be available. Recently, [Ngu16, Ngu15] studied the posterior contraction behavior of several classes of Bayesian hierarchical model (including hierarchical Dirichlet processes) under an analogous setting where the sample size is specified by m sequences of N observations. His approach requires that both m and N tend to infinity and thus cannot be applied to our present setting where N may be fixed.

In this chapter we shall present a parameter estimation theory for general classes of finite mixtures of product distributions. An application of this theory will be posterior contraction theorems established for a standard Bayesian estimation procedure, according to which the de Finetti's mixing measure G tends toward the truth G_0, as m tends to infinity, under suitable conditions. In a standard Bayesian procedure, the statistician endows the space of parameter Θ with a prior distribution Π, which is assumed to have compact support in this theorem, and apply Bayes' rule to obtain the posterior distribution on $\mathcal{E}_{k_0}(\Theta)$, to be denoted by $\Pi(dG|\{X^i_{[N_i]}\}_{i=1}^m)$. To anticipate the distinct convergence behaviors for the atoms and probability mass parameters, for any $G, G' \in \mathcal{E}_k(\Theta)$ defined by

$$D_N(G, G') = \min_{\tau \in S_k} \sum_{i=1}^{k} (\sqrt{N}\|\theta_{\tau(i)} - \theta'_i\|_2 + |p_{\tau(i)} - p'_i|) \quad \forall G = \sum_{i=1}^{k} p_i \delta_{\theta_i}, G' = \sum_{i=1}^{k} p'_i \delta_{\theta'_i} \in \mathcal{E}_k(\Theta),$$

where S_k denotes all the permutations on the set $[k] := \{1, 2, \ldots, k\}$. It can be verified that this is a valid metric in $\mathcal{E}_k(\Theta)$.

Suppose that $N_i = N$ are fixed for all i. We shall naturally require that the sequence length $N \geq n_0$. Moreover, to get fast rate of convergence, we need also $N \geq n_1$ for some minimal natural number $n_1 := n_1(G_0) \geq 1$. We shall call n_1 the *minimal first-order identifiable length*, or 1-identifiable length for short (a formal definition will be given later). In Theorem 2.6.2, it will be established that as soon as $N \geq \max\{n_0, n_1\}$, $\bar{M}_m$ is any sequence of numbers tending to infinity,

the posterior probability

$$\Pi\bigg(G \in \mathcal{E}_{k_0}(\Theta) : D_N(G, G_0) < C(G_0)\bar{M}_m \sqrt{\frac{\ln(mN)}{m}} \bigg| X^1_{[N]}, X^2_{[N]}, \ldots, X^m_{[N]}\bigg)$$

tends to 1 in $\otimes^m P_{G_0,N}$-probability as $m \to \infty$. In the above display, the constant $C(G_0)$ depends on G_0 but is independent of m and N. Expressing this statement more plainly, under the Bayesian estimation procedure, as long as N exceeds 0- and 1-identifiable lengths, the supporting atoms of G converge toward their true values at the rate bounded from above by $(\ln(mN)/(mN))^{1/2}$. Meanwhile, the mixing probabilities p_j converge toward their true values at the rate bounded by $(\ln(mN)/m)^{1/2}$. Note that under additionally stronger identifiability conditions, we may have $n_1 = n_0 = 1$, in which cases it follows that the above claim of posterior contraction rates holds for $N \geq 1$.

In a more realistic setting, for each $i = 1, \ldots, m$, sequence $X^i_{[N_i]}$ is of variable length N_i, which represents the number of repeated measurements for the observed exchangeable sequence. Assume that $\{N_i\}_{i=1}^m$ are uniformly bounded from above by an arbitrary unknown constant, a similar posterior contraction theorem is established (cf. Theorem 2.6.5), where the posterior contraction rate for the mixing proportions remains upper bounded by $m^{-1/2}$, up to a logarithmic quantity. On the other hand, the posterior contraction rate for mixture components' supporting atoms given by

$$O_P\bigg(\sqrt{\frac{\ln(\sum_{i=1}^m N_i)}{\sum_{i=1}^m N_i}}\bigg).$$

Note that the sum $\sum_{i=1}^m N_i$ represents the full volume of the observed data set. Elaborated further, as long as $\min_i N_i \geq \max\{n_0, n_1\}$ and $\sup_i N_i < \infty$, we obtain

$$\Pi\bigg(G \in \mathcal{E}_{k_0}(\Theta) : D_{\sum_{i=1}^m N_i/m}(G, G_0) \leq C(G_0)\bar{M}_m \sqrt{\frac{\ln(\sum_{i=1}^m N_i)}{m}} \bigg| X^1_{[N_1]}, \ldots, X^m_{[N_m]}\bigg) \to 1$$

in $P_{G_0,N_1} \otimes \cdots \otimes P_{G_0,N_m}$-probability as $m \to \infty$. Here, constant $C(G_0)$ is independent of m, sequence lengths $\{N_i\}_{i=1}^m$ and their supremum. In plain terms, we may say that with finite mixtures of product distributions, the posterior inference of atoms of each individual mixture component receives the full benefit of "borrowing strength" across sampled sequences; while the mixing probabilities gain efficiency from only the number of such sequences. This appears to be the first work in which such a posterior contraction theorem is established for de Finetti's mixing measure arising from finite mixtures of product distributions.

The Bayesian learning rates established appear intuitive, given the parameter space $\Theta \in \mathbb{R}^q$ is of finite dimensions. On the role of m, they are somewhat compatible to the previous partial

results [HT00, HZ03, HNPE05]. However, we wish to make several brief remarks at this juncture.

- First, even for exact-fitted parametric mixture models, "parametric-like" learning rates of the form root-m or root-(mN) should not to be taken for granted, because they do not always hold [HN16a, HN19]. This is due to the fact that the kernel family $\{P_\theta\}_{\theta\in\Theta}$ may easily violate assumptions of strong identifiability often required for the root-m rate to take place. In other words, the kernel family $\{P_\theta\}$ may be only *weakly identifiable*, resulting in poor learning rates for a standard mixture, i.e., when $N = 1$ or N is small.

- Second, the fact that by increasing the observed exchangeable sequence's length N so that $N \geq n_1 \vee n_0$, one may obtain parametric-like learning rates in terms of both N and m is a remarkable testament of how repeated measurements can help to completely overcome a latent variable model's potential pathologies, both parameter non-identifiability by making $N \geq n_0$, and parameter estimation inefficiency inherent in a weakly identifiable mixture model, by making $N \geq n_1$. For a deeper appreciation of this issue, we will turn to Section 2.2 for a background on identifiability in parameter estimation as investigated in the literature, which motivates further development in this chapter.

Although the posterior contraction theorems for finite mixtures of product distributions presented in this chapter are new, such results do not adequately capture the complex behavior of the convergence of parameters for a finite mixture of N-product distributions. In fact, the heart of the matter lies in the establishment of a collection of general *inverse bounds*, the type of inequalities of the form

$$D_N(G, G_0) \leq C(G_0)V(P_{G,N}, P_{G_0,N}). \tag{2.2}$$

Note that (2.2) provides an upper bound on distance D_N of mixing measures in terms of the variational distance between the corresponding mixture of N-product distributions. Inequalities of this type allow one to transfer the convergence (and learning rates) of a data population's distribution into that of the corresponding distribution's parameters (therefore the term "inverse bounds"). Several points to highlight are:

- The local nature of (2.2), which may hold only for G residing in a suitably small D_N-neighborhood of G_0 whose radius may also depend on G_0 and N, while constant $C(G_0) > 0$ depends on G_0 but is independent of N. In addition, the bound holds only when N exceeds threshold $n_1 \geq 1$, unless further assumptions are imposed. For instance, under a first-order identifiability condition of P_θ, $n_1 = 1$, so this bound holds for all $N \geq 1$ while remaining local in nature.

- The inverse bounds of the form (2.2) are established without any overt assumption of identifiability. However, they carry striking consequences on both first-order and classical identifia-

bility, where one can deduce from (2.2) under a compactness condition (cf. Proposition 2.5.1): using the notation $n_0(G_0)$ and $n_1(G_0)$ to denote the dependence of 0- and 1-identifiable lengths on G_0, respectively, we have

$$\sup_{G_0 \in \cup_{k \le k_0} \mathcal{E}_k(\Theta^\circ)} n_0(G_0) \le \sup_{G_0 \in \mathcal{E}_{2k_0}(\Theta^\circ)} n_1(G_0) < \infty.$$

Note that classical identifiability captured by n_0 describes a global property of the model family while first-order identifiablity captured by n_1 is local in nature. The connection between these two concepts is made possible because when the number of exchangeable variables N gets large, the force of the central limit theorem comes into play to make the mixture model eventually become identifiable, either in the classical or the first-order sense, even if the model may not be initially identifiable (when $N = 1$).

- The established inverse bounds are sharp in a number of ways. For instance, it can be shown that the quantity N in D_N cannot be improved by $D_{\psi(N)}$ for any sequence $\psi(N)$ such that $\psi(N)/N \to \infty$. Quantifying the effects of identifiability conditions on the strengthening of inverse bounds is a constant theme threading through the chapter.

- These inverse bounds hold for very broad classes of probability kernels $\{P_\theta\}_{\theta \in \Theta}$. In particular, they are established under very mild regularity assumptions on the family of probability kernel P_θ on $\mathfrak{X}$, when either $\mathfrak{X} = \mathbb{R}^d$, or $\mathfrak{X}$ is a finite set, or $\mathfrak{X}$ is an abstract space. A standard but non-trivial example of our theory is the case the kernels P_θ belong to the exponential families of distributions. A more unusual example is the case where P_θ is itself a mixture distribution on $\mathfrak{X} = \mathbb{R}$. Kernels of this type are rarely examined in theory, partly because when we set $N = 1$ a mixture model using such kernels typically would not be parameter-identifiable. However, such "mixture-distribution" kernels are frequently employed by practitioners of hierarchical models (i.e., mixtures of mixture distributions). As the inverse bounds entail, this makes sense since the parameters become strongly identifiable eventually with repeated exchangeable measurements.

- More generally, inverse bounds are established when P_θ does not necessarily admit a density with respect to a dominating measure on $\mathfrak{X}$. An example considered in the chapter is the case P_θ represents probability distribution on the space of probability distributions, namely, P_θ represents (mixtures of) Dirichlet processes. As such, the general inverse bounds are expected to be useful for models with nonparametric mixture components represented by P_θ, the kind of models that have attracted much recent attention, e.g., [TJBB06, RDG08, CLOP19].

The above highlights summarize how the inverse bounds obtained in Section 2.4 and Section 2.5

play the central role in this work. They help to deepen our understanding of the questions of parameter identifiability and provide detailed information about the convergence behavior of parameter estimation. In addition to an asymptotic analysis of Bayesian estimation for mixtures of product distributions that will be carried out in this chapter, such inverse bounds may also be useful for deriving rates of convergence for non-Bayesian parameter estimation procedures, including maximum likelihood estimation and distance based estimation methods. The proofs of these bounds contain novel techniques and insights that may be of independent interest. An overview of the proof will be given in Section 2.2.

The rest of the chapter will proceed as follows. Section 2.2 presents additional related work in the literature and a high-level overview of our approach and techniques. Section 2.3 prepares the reader with basic setups and several useful concepts of distances on space of mixing measures that arise in mixtures of product distributions. Section 2.4 is a self-contained treatment of first-order identifiability theory for finite mixture models, leading to several new results that are useful for subsequent developments. Section 2.5 presents inverse bounds for broad classes of finite mixtures of product distributions, along with specific examples. An immediate application of these bounds are posterior contraction theorems for de Finetti's mixing measures, the main focus of Section 2.6. Section 2.7 gives several technical results demonstrating the sharpness of the established inverse bounds, and which allow to derive minimax lower bounds for estimation procedures of de Finetti's mixing parameters. Particular examples of interest for the inverse bounds established in Section 2.5 include the case the probability kernel P_θ is itself a mixture distribution on $\mathfrak{X} = \mathbb{R}$, and the case P_θ is a mixture of Dirichlet processes. These examples require development of new tools and are deferred to Section 2.8. Finally, (most) proofs of all theorems and lemmas will be provided from Section 2.9 to Section 2.16.

Notation For any probability measure P and Q on measure space $(\mathfrak{X}, \mathcal{A})$ with densities respectively p and q with respect to some base measure μ, the variational distance between them is $V(P,Q) = \sup_{A\in\mathcal{A}} |P(A) - Q(A)| = \int_{\mathfrak{X}} \frac{1}{2}|p(x) - q(x)|d\mu$. The Hellinger distance is given by $h(P,Q) = \left(\int_{\mathfrak{X}} \frac{1}{2}|\sqrt{p(x)} - \sqrt{q(x)}|^2 d\mu\right)^{\frac{1}{2}}$. The Kullback-Leibler divergence of Q from P is $K(p,q) = \int_{\mathfrak{X}} p(x) \ln \frac{p(x)}{q(x)} d\mu$. Note that when a probability measure, say P or P_G, admits a density with respect to a dominating measure, we shall use the lower case to denote such density, say p or p_G, respectively. $\mathcal{B}(\cdot)$ denotes the Borel sigma algebra on the corresponding space. A measurable set $A \in \mathcal{A}$ is μ-positive if $\mu(A) > 0$; is μ-negligible if $\mu(A) = 0$. Write $P \otimes Q$ to be the product measure of P and Q and $\otimes^m P$ for the n-fold product of P.

For any vector $x \in \mathbb{R}^d$, it is a column vector with its i-th coordinate denoted by $x^{(i)}$. The inner product between two vectors a and b is denoted by $a^T b$ or $\langle a, b\rangle$. $\|\cdot\|_2$ for a vector represents its Euclidean distance to the origin. $\|\cdot\|_2$, $\|\cdot\|_F$ for a matrix are respectively its spectral norm and Frobenius norm.

The gradient of a function $f(x)$ at x_0 is $\nabla f(x_0) = \nabla f(x)|_{x=x_0}$. Similar rules apply to partial derivatives or higher order derivatives of a function. The bold $\boldsymbol{i}$ denotes the imaginary number. The maximum of two real numbers a, b is denoted by $\max\{a, b\}$ or $a \vee b$; their minimum is denoted by $\min\{a, b\}$ or $a \wedge b$.

In the presentation of inequality bounds and proofs, $C(\cdot)$, $c(\cdot)$ are positive finite constants depending only on its parameter and may differ from line to line. Write $a \lesssim b$ if $a \leq cb$ for some universal constant c; write $a \lesssim_\xi b$ if $a \leq c(\xi)b$. Write $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$; write $a \asymp_\xi b$ (or $a \overset{\xi}{\asymp} b$) if $a \lesssim_\xi b$ and $b \lesssim_\xi a$.

2.2 Background and overview

2.2.1 First-order identifiability and inverse bounds

In order to shed light on the convergence behavior of model parameters as data sample size increases, stronger forms of identifiability conditions shall be required of the family of probability kernels P_θ. For finite mixture models, such conditions are often stated in terms of a suitable derivative of P_θ with respect to parameter θ, and the linear independence of such derivatives as θ varies in Θ . The impacts of such identifiablity conditions, or the lack thereof, on the convergence of parameter estimation can be quite delicate. Specifically, let $\mathfrak{X} = \mathbb{R}^d$ and fix $N = 1$, so we have $P_G = \sum_{j=1}^k p_j P_{\theta_j}$. Assume that P_θ admits a density function $f(\cdot|\theta)$ with respect to Lebesgue measure on $\mathbb{R}^d$, and for all $x \in \mathbb{R}^d$, $f(\cdot|\theta)$ is differentiable with respect to θ; moreover the combined collection of functions $\{f(\cdot|\theta)\}_{\theta\in\Theta}$ and $\{\nabla f(\cdot|\theta)\}_{\theta\in\Theta}$ are linearly independent. This type of condition, which concerns linear independence of the first derivatives of the likelihood functions with respect to parameter θ, shall be generically referred to as *first-order* identifiability condition of the probability kernel family $\{P_\theta\}_{\theta\in\Theta}$. A version of first-order identifiablity condition was investigated by [HN16b], who showed that their condition will be sufficient for establishing an inverse bound for the form

$$\liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{W_1(G, G_0)} > 0. \tag{2.3}$$

where W_1 denotes the first-order Wasserstein distance metric on $\mathcal{E}_{k_0}(\Theta)$. The infimum limit quantifier should help to clarify somewhat the local nature of the inverse bound (2.2) mentioned earlier. The development of this local inverse bound and its variants plays the fundamental role in the analysis of parameter estimation with finite mixtures in a variety of settings by several authors, where stronger forms of identifiability conditions based on higher order derivatives may be required [Che95, Ngu13, RM11, HN16b, HN16a, HK18, HN19]. In addition, [Ngu13, Ngu16] studied inverse bounds of this type for infinite mixture and hierarchical models.

As noted by [HN16b], for exact-fitted setting of mixtures, i.e., the number of mixture components $k = k_0$ is known, conditions based on only first-order derivatives of P_θ will suffice. Under a suitable first-order identifiability condition based on linear independence of $\{f(\cdot|\theta), \nabla_\theta f(\cdot|\theta)\}_{\theta\in\Theta}$, along with several additional regularity conditions, the mixing measure $G = G_0$ may be estimated via m-i.i.d. sample $(X_1^1, \ldots, X_1^m)$ at the parametric rate of convergence $m^{-1/2}$, due to (2.3) and the fact that the data population density p_{G_0} is typically estimated at the same parametric rate. However, first-order identifiability may not be satisfied, as is the case of two-parameter Gamma kernel, or three-parameter skewnormal kernel, following from the fact that these kernels are governed by certain partial differential equations. In such situations, not only does the resulting Fisher information matrix of the mixture model become singular, the singularity structure of the matrix can be extremely complex — an in-depth treatment of weakly identifiable mixture models can be found in [HN19]. Briefly speaking, in such situations (2.3) may not hold, the rate $m^{-1/2}$ may not be achieved [HN16a, HN19]. In particular, in the case of skewnormal kernels, extremely slow rates of convergence for the component parameters θ_j (e.g., $m^{-1/4}, m^{-1/6}, m^{-1/8}$ and so on) may be established depending on the actual parameter values of the true G_0 for a standard Bayesian estimation or maximum estimation procedure [HN19]. It remains unknown whether it is possible to devise an estimation procedure to achieve the parametric rate of convergence $m^{-1/2}$ when the finite mixture model is only weakly identifiable, i.e., when first-order identifiability condition fails.

In Section 2.4 we shall revisit the described first-order identifiability notions, and then present considerable improvements upon the existing theory and deliver several novel results. First, we identify a tightened set of conditions concerning linear independence of $f(x|\theta)$ and $\nabla_\theta f(x|\theta)$ according to which the inverse bound (2.2) holds. This set of conditions turns out to be substantially weaker than the identifiability condition of [HN16b], most notably by requiring $f(x|\theta)$ be differentiable with respect to θ only for x in a subset of $\mathfrak{X}$ with positive measure. Our weaker notion of first-order identifiability allows us to broaden the scope of probability kernels for which the inverse bound (2.3) continues to apply (cf. Lemma 2.4.2). Second, in a precise sense we show that this notion is in fact necessary for (2.3) to hold (cf. Lemma 2.4.4), giving us an arguably complete characterization of first-order identifiability and its relations to the parametric learning rate for model parameters. Among other new results, it is worth mentioning that when the kernel family $\{P_\theta\}_{\theta\in\Theta}$ belongs to an exponential family of distribution on $\mathfrak{X}$, there is a remarkable equivalence among our notion of first-order identifiability condition and the inverse bound of the form (2.3), and the inverse bound in which variational distance V is replaced by Hellinger distance h (cf. Lemma 2.4.16).

Turning our interest to finite mixtures of product distributions, a key question is on the effect of number N of repeated measurements in overcoming weak identifiability (e.g., the violation of first-order identifiability). One way to formally define the first-order identifiable length (1-identifiable length) $n_1 = n_1(G_0)$ is to make it the minimal natural number such that the following inverse bound

holds for any $N \geq n_1$

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{W_1(G, G_0)} > 0. \tag{2.4}$$

The key question is whether (finite) 1-identifiable length exists, and how can we characterize it. The significance of this concept is that one can achieve first-order identifiability by allowing at least $N \geq n_1$ repeated measurements and obtain the $m^{-1/2}$ learning rate for the mixing measure. In fact, the component parameters can be learned at the rate $(mN)^{-1/2}$, the square root of the full volume of exchangeable data (modulo a logarithmic term). The resolution of the question of existence and characterization of n_1 leads us establish a collection inverse bounds involving mixtures of product distributions that we will describe next. Moreover, such inverse bounds are essential in deriving learning rates for mixing measure G from a collection of exchangeable sequences of observations.

2.2.2 General approach and techniques

For finite mixtures of N-product distributions, for $N \geq 1$, the precise expression for the inverse bound that we aim to establish will be of the form: under certain conditions of the probability kernel $\{P_\theta\}_{\theta \in \Theta}$: for a given $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$, there holds

$$\liminf_{N \to \infty} \liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)} > 0. \tag{2.5}$$

Compared to inverse bound (2.3) for a standard finite mixture, the double infimum limits reveals the challenge for analyzing mixtures of N-product distributions; they express the delicate nature of the inverse bound informally described via (2.2). Moreover, (2.5) entails that n_1 defined by (2.4) exists.

Inverse bound (2.5) will be established for broad classes of kernel P_θ and it can be shown that this bound is sharp. Among the settings of kernel that the bound is applicable, there is a setting when P_θ belongs to any regular exponential family of distributions. More generally, this also include the setting where $\mathfrak{X}$ may be an abstract space and no parametric assumption on P_θ will be required. Instead, we appeal to a set of mild regularity conditions on the characteristic function of a push-forward measure produced by a measurable map T acting on the measure space $(\mathfrak{X}, \mathcal{A})$. Actually an even stronger bound is established relating to the positivity of a suitable notion of curvature on the space of mixtures of product distributions (cf. (2.24)). We will see that this collection of inverse bounds, which are presented in Section 2.5, enables the study for a very broad range of mixtures of product distributions for exchangeable sequences.

The proof of (2.5) and (2.24) represents the core of the chapter. For simplicity, let us first

describe the gist of our proof techniques by considering the case kernel P_θ belongs to an exponential family of distribution on $\mathfrak{X}$ (cf. Theorem 2.5.7). Suppose the kernel admits a density function $p_\theta(x)$ with respect to a dominating measure μ on $\mathfrak{X}$. At a high-level, this is a proof of contradiction: if (2.5) does not hold, then there exists a subsequence $\{N_\ell\}_{\ell=1}^\infty \to \infty$ of natural numbers according to which there exists a sequence of mixing measures $\{G_\ell\}_{\ell=1}^\infty \subset \mathcal{E}_{k_0}(\Theta)\backslash\{G_0\}$ such that $D_{N_\ell}(G_\ell, G_0) \to 0$ as $\ell \to \infty$ and the integral form

$$\frac{V(P_{G_\ell,N_\ell}, P_{G_0,N_\ell})}{D_{N_\ell}(G_\ell, G_0)} = \int_{\mathfrak{X}^{N_\ell}} \left| \frac{p_{G_\ell,N_\ell}(x_1,\ldots,x_{N_\ell}) - p_{G_0,N_\ell}(x_1,\ldots,x_{N_\ell})}{D_{N_\ell}(G_\ell, G_0)} \right| d\otimes^{N_\ell} \mu(x_1,\ldots,x_{N_\ell}) \to 0. \tag{2.6}$$

One may be tempted to applying Fatou's lemma to conclude that the integrand must vanish as $\ell \to \infty$, and from that one may hope to derive an apparent contradiction. This is basically the proof technique of Lemma 2.4.2 for establishing inverse bound (2.3) for finite mixtures, but this would not work here, because the integration domain's dimensionality increases with ℓ. Instead we can exploit the structure of the mixture of N_ℓ-product densities in p_{G_ℓ,N_ℓ}, and rewrite the integral as an expectation with respect to a suitable random variable of fixed domain. What comes to our rescue is the central limit theorem, which is applied to a $\mathbb{R}^q$-valued random variable $Z_\ell = \left(\sum_{n=1}^{N_\ell} T(X_n) - N_\ell \mathbb{E}_{\theta_\alpha^0} T(X_1)\right) / \sqrt{N_\ell}$, where $\mathbb{E}_{\theta_\alpha^0}$ denotes the expectation taken with respect to the probability distribution P_θ for some suitable $\theta = \theta_\alpha^0$ chosen among the support of true mixing measure G_0. Here $T : \mathfrak{X} \to \mathbb{R}^q$ denotes the sufficient statistic for the exponential family distribution $P_\theta(dx_n)$, for each $n = 1, \ldots, N_\ell$.

Continuing with this plan, by a change of measure the integral in (2.6) may be expressed as the expectation of the form $\mathbb{E}|\Psi_\ell(Z_\ell)|$ for some suitable function $\Psi_\ell : \mathbb{R}^q \to \mathbb{R}$. By exploiting the structure of the exponential families dictating the form of Ψ_ℓ, it is possible to obtain that for any sequence $z_\ell \to z$, there holds $\Psi_\ell(z_\ell) \to \Psi(z)$ for a certain function $\Psi : \mathbb{R}^q \to \mathbb{R}$. Since Z_ℓ converges in distribution to Z a non-degenerate zero-mean Gaussian random vector in $\mathbb{R}^q$, it entails that $\Psi_\ell(Z_\ell)$ converges to $\Psi(Z)$ in distribution by a generalized continuous mapping theorem [WVdV96]. Coupled with a generalized Fatou's lemma [Bil96], we arrive at $\mathbb{E}|\Psi(Z)| = 0$, which can be verified as a contradiction.

For the general setting where $\{P_\theta\}_{\theta\in\Theta}$ is a family of probability on measure space $(\mathfrak{X}, \mathcal{A})$, the basic proof structure remains the same, but we can no longer exploit the parametric assumption on the kernel family P_θ (cf. Theorem 2.5.14). Since the primary object of inference is parameter $\theta \in \Theta \subset \mathbb{R}^q$, the assumptions on the kernel P_θ will center on the existence of a measurable map $T : (\mathfrak{X}, \mathcal{A}) \to (\mathbb{R}^s, \mathcal{B}(\mathbb{R}^s))$ for some $s \geq q$, and regularity conditions on the push-forward measure on $\mathbb{R}^s$: $P_\theta \# T := P_\theta \circ T^{-1}$. This measurable map plays the same role as that of sufficient statistic T when P_θ belongs to the exponential family. The main challenge lies in the analysis of function Ψ_ℓ described in the previous paragraph. It is here that the power of Fourier analysis is brought

to bear on the analysis of Ψ_ℓ and the expectation $\mathbb{E}_{\theta^0_\alpha}\Psi_\ell(Z_\ell)$. By the Fourier inversion theorem, Ψ_ℓ may be expressed entirely in terms of the characteristic function of the push-forward measure $P_\theta \# T$. Provided regularity conditions on such characteristic function hold, one is able to establish the convergence of Ψ_ℓ toward a certain function $\Psi : \mathbb{R}^s \to \mathbb{R}$ as before.

We shall provide a variety of examples demonstrating the broad applicability of Theorem 2.5.14, focusing on the cases P_θ does not belong to an exponential family of distributions. In some cases, checking for the existence of map T is straightforward. When P_θ is a complex object, in particular, when P_θ is itself a mixture distribution, this requires substantial work, as should be expected. In this example, the burden of checking the applicability of Theorem 2.5.14 lies primarily in evaluating certain oscillatory integrals composed of the map T in question. Tools from harmonic analysis of oscillatory integrals will be developed for such a purpose and presented in Section 2.8. We hope that the tools developed here present a useful stepping stone toward a more satisfactory asymptotic treatment of complex hierarchical models (models that may be viewed as mixtures of mixtures of distributions, e.g. [TJBB06, RDG08, Ngu16, CLOP19]), which have received broad and increasingly deepened attention in the literature.

2.3 Preliminaries

We shall start by setting up basic notions required for the analysis of mixtures of product distributions. Throughout this chapter the exchangeable data sequences are denoted by $X^i_{[N_i]} := (X^i_1, \ldots, X^i_{N_i})$ for $i = 1, \ldots, m$, while N_i denotes the length of sequence $X^i_{[N_i]}$. For ease of presentation, for now, we shall assume that $N_i = N$ for all i. Later on we will allow the observed exchangeable sequences to be of variable lengths. These sequences are composed of elements in a measurable space $(\mathfrak{X}, \mathcal{A})$. Examples include $\mathfrak{X} = \mathbb{R}^d$, $\mathfrak{X}$ is a discrete space, $\mathfrak{X}$ is a space of measures. Regardless, the parameters of interest are always encapsulated by a discrete mixing measure $G \in \mathcal{E}_k(\Theta)$, the space of discrete measures with k distinct support atoms residing in a set $\Theta \subset \mathbb{R}^q$.

The linkage between parameters of interest, i.e., the mixing measure G, and the observed data sequences is achieved via the mixture of product distributions that we now define. Consider a family of probability distributions $\{P_\theta\}_{\theta\in\Theta}$ on measurable space $(\mathfrak{X}, \mathcal{A})$, where θ is the parameter of the family and $\Theta \subset \mathbb{R}^q$ is the parameter space. For $N \in \mathbb{N}$, the N-product probability family is denoted by $\{P_{\theta,N} := \bigotimes^N P_\theta\}_{\theta\in\Theta}$ on $(\mathfrak{X}^N, \mathcal{A}^N)$, where $\mathcal{A}^N$ is the product sigma algebra. Given a mixing measure $G = \sum_{i=1}^k p_i \delta_{\theta_i} \in \mathcal{E}_k(\Theta)$, the mixture of N-product distributions induced by G is given by

$$P_{G,N} = \sum_{i=1}^k p_i P_{\theta_i,N}.$$

Each exchangeable sequence $X^i_{[N]} = (X^i_1, \ldots, X^i_N)$, for $i = 1, \ldots, m$, is an independent sample distributed according to $P_{G,N}$. Due to the role they play in the composition of distribution $P_{G,N}$, we also refer to $\{P_\theta\}_{\theta\in\Theta}$ as a family of *probability kernels* on $(\mathfrak{X}, \mathcal{A})$.

In order to quantify the convergence of mixing measures arising in mixture models, an useful device is a suitably defined optimal transport distance [Ngu13, Ngu11]. Consider the Wasserstein-p distance w.r.t. distance d_θ on Θ: $\forall G = \sum_{i=1}^k p_i\delta_{\theta_i}, G' = \sum_{i=1}^{k'} p'_i\delta_{\theta'_i}$, define

$$W_p(G, G'; d_\theta) = \left(\min_{\boldsymbol{q}} \sum_{i=1}^{k}\sum_{j=1}^{k'} q_{ij} d^p_\theta(\theta_i, \theta'_j)\right)^{1/p}, \tag{2.7}$$

where the infimum is taken over all joint probability distributions $\boldsymbol{q}$ on $[k] \times [k']$ such that, when expressing $\boldsymbol{q}$ as a $k \times k'$ matrix, the marginal constraints hold: $\sum_{j=1}^{k'} q_{ij} = p_i$ and $\sum_{i=1}^{k} q_{ij} = p'_j$. For the special case when d_θ is the Euclidean distance, write simply $W_p(G, G')$ instead of $W_p(G, G'; d_\theta)$. Write $G_\ell \overset{W_p}{\to} G$ if G_ℓ converges to G under the W_p distance w.r.t. the Euclidean distance on Θ.

We will see in this chapter that for mixing measures arising in mixtures of N-product distributions, a more useful notion is the following. For any $G = \sum_{i=1}^k p_i\delta_{\theta_i} \in \mathcal{E}_k(\Theta)$ and $G' = \sum_{i=1}^k p'_i\delta_{\theta'_i} \in \mathcal{E}_k(\Theta)$, define

$$D_N(G, G') = \min_{\tau\in S_k} \sum_{i=1}^{k} (\sqrt{N}\|\theta_{\tau(i)} - \theta'_i\|_2 + |p_{\tau(i)} - p'_i|)$$

where S_k denote all the permutations on the set $[k]$.

It is simple to verify that $D_N(\cdot, \cdot)$ is a valid metric on $\mathcal{E}_k(\Theta)$ for each N and relate it to a suitable optimal transport distance metric. Indeed, $G = \sum_{i=1}^k p_i\delta_{\theta_i} \in \mathcal{E}_k(\Theta)$, due to the permutations invariance of its atoms, can be identified as a set $\{(\theta_i, p_i) : 1 \le i \le k\}$, which can further be identified as $\tilde{G} = \sum_{i=1}^k \frac{1}{k}\delta_{(\theta_i,p_i)} \in \mathcal{E}_k(\Theta \times \mathbb{R})$. Formally, we define a map $\mathcal{E}_k(\Theta) \to \mathcal{E}_k(\Theta \times \mathbb{R})$ by

$$G = \sum_{i=1}^{k} p_i\delta_{\theta_i} \mapsto \tilde{G} = \sum_{i=1}^{k} \frac{1}{k}\delta_{(\theta_i,p_i)} \in \mathcal{E}_k(\Theta \times \mathbb{R}). \tag{2.8}$$

Now, endow a metric M_N on $\Theta \times \mathbb{R}$ defined by $M_N((\theta, p), (\theta', p')) = \sqrt{N}\|\theta - \theta'\|_2 + |p - p'|$ and note the following fact.

Lemma 2.3.1. *For any $G = \sum_{i=1}^k \frac{1}{k}\delta_{\theta_i}, G' = \sum_{i=1}^k \frac{1}{k}\delta_{\theta'_i} \in \mathcal{E}_k(\Theta)$ and any distance d_θ on Θ,*

$$W^p_p(G, G'; d_\theta) = \min_{\tau\in S_k} \frac{1}{k}\sum_{i=1}^{k} d^p_\theta(\theta_i, \theta'_{\tau(i)}).$$

A proof of the preceding lemma is available as Proposition 2 in [Ngu11]. Apply Lemma 2.3.1 with Θ, d_θ replaced respectively by $\Theta \times \mathbb{R}$ and M_N, then for any $G, G' \in \mathcal{E}_k(\Theta)$, $W_1(\tilde{G}, \tilde{G}'; M_N) = \frac{1}{k} D_N(G, G')$, which validates that D_N is indeed a distance on $\mathcal{E}_k(\Theta)$, and moreover it does not depend on the specific representations of G and G'.

The next lemma establishes the relationship between D_N and W_1 on $\mathcal{E}_k(\Theta)$.

Lemma 2.3.2. *a) A sequence $G_n \in \mathcal{E}_k(\Theta)$ converges to $G_0 \in \mathcal{E}_k(\Theta)$ under W_p if and only if G_n converges to G_0 under D_N. That is, W_p and D_N generate the same topology.*

b) Let Θ be bounded. Then $W_1(G, G') \leq \max\left\{1, \frac{diam(\Theta)}{2}\right\} D_1(G, G')$ for any $G, G' \in \mathcal{E}_k(\Theta)$. More generally for any $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^k p'_i \delta_{\theta'_i}$,

$$W_p^p(G, G') \leq \max\left\{1, \frac{diam^p(\Theta)}{2}\right\} \min_{\tau \in S_k} \sum_{i=1}^k \left(\|\theta_{\tau(i)} - \theta'_i\|_2^p + |p_{\tau(i)} - p'_i| \right).$$

c) Suppose Θ° is not empty. Then $\inf_{G, G' \in \mathcal{E}_k(\Theta)} \frac{W_1(G,G')}{D_1(G,G')} = 0$.

d) Fix $G_0 \in \mathcal{E}_k(\Theta)$. Then $\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_k(\Theta)}} \frac{W_1(G,G_0)}{D_1(G,G_0)} > 0$ and $\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_k(\Theta)}} \frac{D_1(G,G_0)}{W_1(G,G_0)} > 0$. That is, in a neighborhood of G_0 in $\mathcal{E}_k(\Theta)$, $D_1(G, G_0) \asymp_{G_0} W_1(G, G_0)$.

e) Fix $G_0 \in \mathcal{E}_k(\Theta)$ and suppose Θ is bounded. Then $W_1(G, G_0) \geq C(G_0, diam(\Theta)) D_1(G, G_0)$ for any $G \in \mathcal{E}_k(\Theta)$, where $C(G_0, diam(\Theta))$ is a constant that depends on G_0 and $diam(\Theta)$.

See Section 2.9 for a proof. We see that W_1 and D_1 generate the same topology on $\mathcal{E}_k(\Theta)$, and they are equivalent in the sense that they differ from each other by only a constant factor while fixing one argument. The benefit of W_p is that it is defined on $\bigcup_{k=1}^\infty \mathcal{E}_k(\Theta)$ while D_N is only defined $\mathcal{E}_k(\Theta)$ for each k since its definition requires the two arguments has the same number of atoms. The benefit of using D_N is that it allows us to quantify the distinct convergence behavior for atoms and probability mass, by placing different factors (or even different powers, cf.(2.17)) on the atoms and the probability mass, while W_p fails to do so. For example, one may tempt to consider $W_p^p(G, G'; d_\theta)$ with $d_\theta(\theta, \theta') = \sqrt{N}\|\theta - \theta'\|_2$ to put a $\sqrt{N}$ factor on the atoms but it will not work as illustrated by the following example.

Example 2.3.3. Consider $G_1 = p_1^1 \delta_{\theta_1} + p_2^1 \delta_{\theta_2}$ and $G_2 = p_1^2 \delta_{\theta_1} + p_2^2 \delta_{\theta_2} \in \mathcal{E}_2(\Theta)$ with $p_1^1 \neq p_1^2$. When N is sufficiently large, $D_N(G_1, G_2) = |p_1^1 - p_1^2| + |p_2^1 - p_2^2|$, a constant independent of N. But with d_θ being Euclidean distance multiplied by $\sqrt{N}$

$$W_p^p(G_1, G_2; d_\theta)$$

$$= \min_{\boldsymbol{q}} (q_{12} + q_{21}) \left(\sqrt{N}\|\theta_1 - \theta_2\|_2\right)^p = \left(\sqrt{N}\|\theta_1 - \theta_2\|_2\right)^p \frac{1}{2}\left(|p_1^1 - p_1^2| + |p_2^1 - p_2^2|\right),$$

where $\boldsymbol{q}$ is a coupling as in (2.7). So $W_p(G_1, G_2; d_\theta) = \sqrt{N}\|\theta_1 - \theta_2\|_2 \left(\frac{1}{2}(|p_1^1 - p_1^2| + |p_2^1 - p_2^2|)\right)^{1/p}$, which increases to ∞ when $N \to \infty$. Even G_1 and G_2 has the set of atoms, $W_p(G_1, G_2; d_\theta)$ is still of the order $\sqrt{N}$. Thus, $W_p(G_1, G_2; d_\theta)$ couple atoms and probability; in other words it does not separate them in the way D_N does. In the sequel it will be shown that D_N is the "right" distance to use when we develop general inverse bounds of mixtures of product measures. □

2.4 First-order identifiability theory

Let $N = 1$, a finite mixture of N-product distributions is reduced to a standard finite mixture of distributions. Mixture components are modeled by a family of probability kernels $\{P_\theta\}_{\theta\in\Theta}$ on $\mathfrak{X}$, where θ is the parameter of the family and $\Theta \subset \mathbb{R}^q$ is the parameter space. As discussed in the introduction, throughout the chapter we assume that the map $\theta \mapsto P_\theta$ is injective; it is the nature of the map $G \mapsto P_G$ that we are after. Within this section, we further assume that $\{P_\theta\}_{\theta\in\Theta}$ has density $\{f(x|\theta)\}_{\theta\in\Theta}$ w.r.t. a dominating measure μ on $(\mathfrak{X}, \mathcal{A})$. Combining multiple mixture components using a mixing measure G on Θ results in the finite mixture distribution, which admits the following density with respect to μ: $p_G(x) = \int f(x|\theta)G(d\theta)$. The goal of this section is to provide a concise and self-contained treatment of identifiability of finite mixture models. We lay down basic foundations and present new results that will prove useful for the general theory of mixtures of product distributions to be developed in the subsequent sections.

2.4.1 Basic theory

The classical identifiability condition posits that P_G uniquely identify G for all $G \in \mathcal{E}_{k_0}(\Theta)$. This condition is satisfied if the collection of density functions $\{f(x|\theta)\}_{\theta\in\Theta}$ are linearly independent. In order to obtain rates of convergence for the model parameters, it is natural to consider the following condition concerning the first-order derivative of f with respect to θ.

Definition 2.4.1. The family $\{f(x|\theta)\}_{\theta\in\Theta}$ is $(\{\theta_i\}_{i=1}^k, \mathcal{N})$ **first-order identifiable** if

(i) for every x in the μ-positive subset $\mathfrak{X}\backslash\mathcal{N}$ where $\mathcal{N} \in \mathcal{A}$, $f(x|\theta)$ is first-order differentiable w.r.t. θ at $\{\theta_i\}_{i=1}^k$; and

(ii) $\{\theta_i\}_{i=1}^k \subset \Theta^\circ$ is a set of k distinct elements and the system of two equations with variable $(a_1, b_1, \ldots, a_k, b_k)$:

$$\sum_{i=1}^k \left(a_i^T \nabla_\theta f(x|\theta_i) + b_i f(x|\theta_i)\right) = 0, \quad \mu - a.e.\ x \in \mathfrak{X}\backslash\mathcal{N}, \tag{2.9a}$$

$$\sum_{i=1}^{k} b_i = 0 \tag{2.9b}$$

has only the zero solution: $b_i = 0 \in \mathbb{R}$ and $a_i = \mathbf{0} \in \mathbb{R}^q, \quad \forall 1 \leq i \leq k$.

This definition specifies a condition that is weaker than the definition of *identifiable in the first-order* in [HN16b] since it only requires $f(x|\theta)$ to be differentiable at a finite number of points $\{\theta_i\}_{i=1}^k$. Moreover, it does *not* require $f(x|\theta)$ as a function of θ to be differentiable for μ-a.e. x. Our defined first-order identifiability requires only linear independence between the density and its derivative w.r.t. the parameter over the constraints of the coefficients specified by (2.9b). We will see shortly that in a precise sense that the conditions given Definition 2.4.1 are also necessary.

The significance of first-order identifiability conditions is that they entail a collection of inverse bounds that relate the behavior of some form of distances on mixture densities P_G, P_{G_0} to a distance between corresponding parameters described by $D_1(G, G_0)$, as G tends toward G_0. For any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$, and define

$$B_{W_1}(G_0, r) = \left\{ G \in \bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta) \middle| W_1(G, G_0) < r \right\}. \tag{2.10}$$

It's obvious that $B_{W_1}(G_0, r) \subset \mathcal{E}_{k_0}(\Theta)$ for small r.

Lemma 2.4.2 (Consequence of first-order identifiability)**.** *Let $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$. Suppose that the family $\{f(x|\theta)\}_{\theta\in\Theta}$ is $(\{\theta_i^0\}_{i=1}^{k_0}, \mathcal{N})$ first-order identifiable in the sense of Definition 2.4.1 for some $\mathcal{N} \in \mathcal{A}$.*

a) Then

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{D_1(G, G_0)} > 0. \tag{2.11}$$

b) If in addition, for every x in $\mathfrak{X}\backslash\mathcal{N}$ $f(x|\theta)$ is continuously differentiable w.r.t. θ in a neighborhood of θ_i^0 for $i \in [k_0]$, then

$$\lim_{r\to 0} \inf_{\substack{G,H\in B_{W_1}(G_0,r) \\ G\neq H}} \frac{V(P_G, P_H)}{D_1(G, H)} > 0. \tag{2.12}$$

To put the above claims in context, note that the following inequality holds generally for *any* probability kernel family $\{P_\theta\}_{\theta\in\Theta}$ (even those without a density w.r.t. a dominating measure), cf.

Lemma 2.7.1:

$$\sup_{G_0\in\mathcal{E}_{k_0}(\Theta)} \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G,P_{G_0})}{D_1(G,G_0)} \leq 1/2. \tag{2.13}$$

Note also that

$$\lim_{r\to 0} \inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}} \frac{V(P_G,P_H)}{D_1(G,H)} \leq \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G,P_{G_0})}{D_1(G,G_0)}$$

for any probability kernel P_θ and for any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$. Then (2.12) immediately entails (2.11). However, (2.11) is sufficient for translating a learning rate for estimating a population distribution P_G into that of the corresponding mixing measure G. To be concrete, if we are given an m-i.i.d. sample from a parametric model P_{G_0}, a standard estimation method would yield root-m rate of convergence for density p_G, which means that the corresponding estimate of G admits root-m rate as well.

Remark 2.4.3. Lemma 2.4.2 is a generalization of the Theorem 3.1 in [HN16b] in several key features. Firstly, $(\{\theta_i^0\}_{i=1}^{k_0}, \mathcal{N})$ first-order identifiable assumption in Lemma 2.4.2 is weaker since any identifiable in the first-order in the sense of [HN16b] implies $(\{\theta_i^0\}_{i=1}^{k_0}, \mathcal{N})$ first-order identifiable with $\mathcal{N} = \emptyset$. Example 2.10.1 gives a specific instance which satisfies the notion of first-order identifiability specified by Definition 2.4.1 but not the condition specified by [HN16b]. Secondly, it turns out that uniform Lipschitz assumption in Theorem 3.1 in [HN16b] is redundant and Lemma 2.4.2 does not require it (such Lipschitz condition appears to be still needed for the inverse bounds on the sup-norm of mixture densities, cf. [HN16b]). Finally, given some additional features of f, the first-order identifiable notion can be further simplified; such details will be given in Section 2.4.2. □

Proof of Lemma 2.4.2 a): Suppose the lower bound of (2.11) is incorrect. Then there exist $G_\ell \in \mathcal{E}_{k_0}(\Theta)\backslash\{G_0\}$, $G_\ell \overset{W_1}{\to} G_0$ such that

$$\frac{V(p_{G_\ell},p_{G_0})}{D_1(G_\ell,G_0)} \to 0, \text{ as } \ell\to\infty.$$

We may write $G_\ell = \sum_{i=1}^{k_0} p_i^\ell \delta_{\theta_i^\ell}$ such that $\theta_i^\ell \to \theta_i^0$ and $p_i^\ell \to p_i^0$ as $\ell\to\infty$. With subsequences argument if necessary, we may futher require

$$\frac{\theta_i^\ell-\theta_i^0}{D_1(G_\ell,G_0)} \to a_i \in \mathbb{R}^q, \quad \frac{p_i^\ell - p_i^0}{D_1(G_\ell,G_0)} \to b_i \in \mathbb{R}, \quad \forall 1\leq i\leq k_0, \tag{2.14}$$

where b_i and the components of a_i are in $[-1,1]$ and $\sum_{i=1}^{k_0} b_i = 0$. Moreover, $D_1(G_\ell,G_0) =$

$\sum_{i=1}^{k_0} \left(\|\theta_i^\ell - \theta_i^0\|_2 + |p_i^\ell - p_i^0| \right)$ for sufficiently large ℓ, which implies

$$\sum_{i=1}^{k_0} \|a_i\|_2 + \sum_{i=1}^{k_0} |b_i| = 1.$$

It also follows that at least one of a_i is not $\mathbf{0} \in \mathbb{R}^q$ or one of b_i is not 0. On the other hand,

$$\begin{aligned}
0 &= \lim_{\ell\to\infty} \frac{2V(P_{G_\ell}, P_{G_0})}{D_1(G_\ell, G_0)} \\
&\geq \lim_{\ell\to\infty} \int_{\mathfrak{X}\setminus\mathcal{N}} \left| \sum_{i=1}^{k_0} p_i^\ell \frac{f(x|\theta_i^\ell) - f(x|\theta_i^0)}{D_1(G_\ell, G_0)} + \sum_{i=1}^{k_0} f(x|\theta_i^0) \frac{p_i^\ell - p_i^0}{D_1(G_\ell, G_0)} \right| \mu(dx) \\
&\geq \int_{\mathfrak{X}\setminus\mathcal{N}} \liminf_{\ell\to\infty} \left| \sum_{i=1}^{k_0} p_i^\ell \frac{f(x|\theta_i^\ell) - f(x|\theta_i^0)}{D_1(G_\ell, G_0)} + \sum_{i=1}^{k_0} f(x|\theta_i^0) \frac{p_i^\ell - p_i^0}{D_1(G_\ell, G_0)} \right| \mu(dx) \\
&= \int_{\mathfrak{X}\setminus\mathcal{N}} \left| \sum_{i=1}^{k_0} p_i^0 a_i^T \nabla_\theta f(x|\theta_i^0) + \sum_{i=1}^{k_0} f(x|\theta_i^0) b_i \right| \mu(dx).
\end{aligned}$$

where the second inequality follows from Fatou's Lemma. Then

$$\sum_{i=1}^{k_0} p_i^0 a_i^T \nabla_\theta f(x|\theta_i^0) + \sum_{i=1}^{k_0} f(x|\theta_i^0) b_i = 0$$

for $\mu - a.e.x \in \mathfrak{X}\setminus\mathcal{N}$. Thus we find a nonzero solution to (2.9a), (2.9b) with k, θ_i replaced by k_0, θ_i^0.

However, the last statement contradicts with the definition of $(\{\theta_i^0\}_{i=1}^{k_0}, \mathcal{N})$ first-order identifiable.

Proof of part b) continues in the Appendix. □

Lemma 2.4.2 states that under (i) in Definition 2.4.1, the constrained linear independence between the density and its derivative w.r.t. the parameter (item (ii) in the definition) is sufficient for (2.11) and (2.12). For a converse result, the next lemma shows (ii) is also necessary provided that (i) holds for some μ-negligible $\mathcal{N}$ and the density $f(x|\theta)$ satisfies some regularity condition.

Lemma 2.4.4 (Lack of first-order identifiability)**.** *Fix $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$. Suppose*

a) there exists $\mathcal{N}$ (that possibly depends on G_0) such that $\mu(\mathcal{N}) = 0$ and for every $x \notin \mathcal{N}$, $f(x|\theta)$ is differentiable w.r.t. θ at $\{\theta_i^0\}_{i=1}^{k_0}$;

b) equation (2.9a) *(or equivalently, system of equations* (2.9a) *and* (2.9b)*) with k, θ_i replaced respectively by k_0, θ_i^0 has a nonzero solution $(a_1, b_1, \ldots, a_{k_0}, b_{k_0})$;*

c) there exist $\gamma_0 > 0$ *such that* $\forall\ 1 \le i \le k_0$, $\forall\ 0 < \Delta \le \gamma_0$,

$$\left| \frac{f(x|\theta_i^0 + a_i\Delta) - f(x|\theta_i^0)}{\Delta} \right| \le \bar{f}(x), \quad \mu - a.e.\ x \in \mathfrak{X}\backslash\mathcal{N},$$

where $\bar{f}(x)$ *is integrable w.r.t. the measure* μ*;*

then

$$\lim_{r\to 0} \inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}} \frac{V(P_G, P_H)}{D_1(G,H)} = \liminf_{\substack{G\xrightarrow{W_1} G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{D_1(G,G_0)} = 0. \tag{2.15}$$

Remark 2.4.5. Condition c) in the preceding lemma is to guarantee the exchange of the order between limit and the integral and one may replace it by any other similar condition. A byproduct of this condition is that it renders the constraint (2.9b) redundant (cf. Lemma 2.4.14 b)). While condition c) is tailored for an application of the dominated convergence theorem in the proof, one may tailored the following condition for Pratt's Lemma.

Condition c'): there exist $\gamma_0 > 0$ such that $\forall\ 1 \le i \le k_0$, $\forall\ 0 < \Delta < \gamma_0$,

$$\left| \frac{f(x|\theta_i^0 + a_i\Delta) - f(x|\theta_i^0)}{\Delta} \right| \le \bar{f}_\Delta(x), \quad \mu - a.e.\ x \in \mathfrak{X}\backslash\mathcal{N}$$

where $\bar{f}_\Delta(x)$ satisfies $\lim_{\Delta\to 0^+} \int_{\mathfrak{X}\backslash\mathcal{N}} \bar{f}_\Delta(x)d\mu = \int_{\mathfrak{X}\backslash\mathcal{N}} \lim_{\Delta\to 0^+} \bar{f}_\Delta(x)d\mu$.

Condition c') is weaker than condition c) since the former reduces to the latter if one let $\bar{f}_\Delta(x) = \bar{f}(x) < \infty$. □

Combining all the conditions in Lemma 2.4.2 and Lemma 2.4.4, one immediately the following equivalence between (2.11), (2.12) and the first-order identifiable condition.

Corollary 2.4.6. *Fix* $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$. *Suppose for* μ*-a.e.* $x \in \mathfrak{X}$, $f(x|\theta)$ *as a function* θ *is continuously differentiable in a neighborhood of* θ_i^0 *for each* $i \in [k_0]$. *Suppose that for any* $a \in \mathbb{R}^q$ *and for each* $i \in [k_0]$ *there exists* $\gamma(\theta_i^0, a) > 0$ *such that for any* $0 < \Delta \le \gamma(\theta_i^0, a)$,

$$\left| \frac{f(x|\theta_i^0 + a\Delta) - f(x|\theta_i^0)}{\Delta} \right| \le \bar{f}_\Delta(x|\theta_i^0, a) \quad \mu - a.e.\ \mathfrak{X} \tag{2.16}$$

where $\bar{f}_\Delta(x|\theta_i^0, a)$ *satifies*

$$\lim_{\Delta\to 0^+} \int_{\mathfrak{X}} \bar{f}_\Delta(x|\theta_i^0, a)d\mu = \int_{\mathfrak{X}} \lim_{\Delta\to 0^+} \bar{f}_\Delta(x|\theta_i^0, a)d\mu.$$

Here $\bar{f}_\Delta(x|\theta_i, a_i)$ *possibly depends on* θ_i^0 *and* a. *Then* (2.12) *holds if and only if* (2.11) *holds if and only if* (2.9a) *with* k, θ_i *replaced respectively by* k_0, θ_i^0 *has only the zero solution.*

The Lemma 2.4.4 presents the consequence of the violation of first-order identifiability. Indeed, the conclusion (2.15) suggests that $D_1(G, G_0)$ may vanish at a much slower rate than $V(P_G, P_{G_0})$, i.e., the convergence of parameters representing G may be much slower than the convergence of data distribution P_G. Moreover, the impact may be different for different types of parameters, as already noted in [HN19]. To explicate the distinct behavior for the mixing probability parameters p_i and for the atoms θ_i, define the following notion of distance, for any $G = \sum_{i=1}^k p_i \delta_{\theta_i} \in \mathcal{E}_k(\Theta)$ and $G = \sum_{i=1}^k p'_i \delta_{\theta'_i} \in \mathcal{E}_k(\Theta)$, $r_1 \geq 1$ and $r_2 \geq 1$,

$$D_{r_1,r_2}(G, G') := \min_{\tau \in S_k} \sum_{i=1}^k \left(\|\theta_i - \theta'_{\tau(i)}\|_2^{r_1} + |p_i - p'_{\tau(i)}|^{r_2} \right). \tag{2.17}$$

Although D_{r_1,r_2} might not be a proper metric, but by Lemma 2.3.2 b) and similar to Lemma 2.3.2 e), for a fixed $G_0 \in \mathcal{E}_k(\Theta)$, $D_{r,1}(G_0, G) \asymp_{G_0, \text{diam}(\Theta)} W_r^r(G_0, G)$ provided that Θ is bounded, i.e. there exists $c_1(G_0, \text{diam}(\Theta))$ and $c_2(\text{diam}(\Theta))$ such that for any $G \in \mathcal{E}_k(\Theta)$

$$c_1(G_0, \text{diam}(\Theta)) D_{r,1}(G_0, G) \leq W_r^r(G_0, G) \leq c_2(\text{diam}(\Theta)) D_{r,1}(G_0, G).$$

The conclusion of Lemma 2.4.4 can be refined further by the following result.

Lemma 2.4.7 (Impacts on different parameters)**.** *Suppose all conditions in Lemma 2.4.4 are satisfied.*

a) If at least one of b_i is not zero, then for any $r \geq 1$

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{D_{r,1}(G, G_0)} = \liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{W_r^r(G, G_0)} = 0.$$

b) If at least one of a_i is not zero, then for any $r \geq 1$

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{D_{1,r}(G, G_0)} = 0.$$

Next, we highlight the role of condition c) of Lemma 2.4.4 in establishing either inverse bound (2.11) or (2.15) based on our notion of first-order identifiability. As mentioned, condition c) posits the existence of an integrable envelope function to ensure the exchange of the limit and integral. Without this condition, the conclusion (2.15) of Lemma 2.4.4 might not hold. The following two examples demonstrate the role of c), and serve as examples which are not first-order identifiable but for which inverse bound (2.11) still holds.

Example 2.4.8 (Uniform probability kernel)**.** Consider the uniform distribution family $f(x|\theta) = \frac{1}{\theta}\mathbf{1}_{(0,\theta)}(x)$ with parameter space $\Theta = (0,\infty)$. This family is defined on $\mathfrak{X} = \mathbb{R}$ with the dominating measure μ to be the Lebesgue measure. It is easy to see $f(x|\theta)$ is differentiable w.r.t. θ at $\theta \neq x$ and

$$\frac{\partial}{\partial\theta} f(x|\theta) = -\frac{1}{\theta} f(x|\theta) \quad \text{when } \theta \neq x.$$

So $f(x|\theta)$ is not first-order identifiable by our definition. Note for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$ this family does not satisfy the assumption c) in Lemma 2.4.4 and hence Lemma 2.4.4 is not applicable. Indeed by Lemma 2.4.9 this family satisfies (2.11) and (2.12) for any k_0 and $G_0 \in \mathcal{E}_{k_0}(\Theta)$. □

Lemma 2.4.9. *Let $f(x|\theta)$ be the uniform distribution family defined in Example 2.4.8. Then for any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$, inverse bounds* (2.11) *and* (2.12) *hold.*

Example 2.4.10 (Location-scale exponential distribution kernel)**.** Consider the location-scale exponential distribution on $\mathfrak{X} = \mathbb{R}$, with density with respect to μ =Lebesgue measure given by $f(x|\xi,\sigma) = \frac{1}{\sigma}\exp\left(-\frac{x-\xi}{\sigma}\right)\mathbf{1}_{(\xi,\infty)}(x)$ with parameter $\theta = (\xi,\sigma)$ and parameter space $\Theta = \mathbb{R}\times(0,\infty)$. It is easy to see $f(x|\xi,\sigma)$ is differentiable w.r.t. ξ at $\xi \neq x$ and

$$\frac{\partial}{\partial\xi} f(x|\xi,\sigma) = \frac{1}{\sigma} f(x|\xi,\sigma) \quad \text{when } \xi \neq x.$$

So $f(x|\xi,\sigma)$ is not first-order identifiable. Note for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$ this family does not satisfy the third assumption in Lemma 2.4.4 and hence Lemma 2.4.4 is not applicable. Indeed by Lemma 2.4.11 this family satisfies (2.11) for any k_0 and $G_0 \in \mathcal{E}_{k_0}(\Theta)$. This lemma also serves as a correction for an errornous result (Prop. 5.3 of [HN16a]). The mistake in their proof may be attributed to failing to account for the envelope condition c) that arises due to non-identical support of mixture components with distinct ξ values. □

Lemma 2.4.11. *Let $f(x|\xi,\sigma)$ be the location-scale exponential distribution defined in Example 2.4.10. Then for any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$, inverse bound* (2.11) *holds.*

In some context it is of interest of establish inverse bounds for Hellinger distance rather than variational distance on mixture densities, e.g., in the derivation of minimax lower bounds. Since $h \geq V$, the inverse bound (2.11), which holds under first-order identifiability, immediately entails that

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{h(P_G, P_{G_0})}{D_1(G, G_0)} > 0. \tag{2.18}$$

Similarly, (2.12) entails that

$$\lim_{r\to 0}\inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}}\frac{h(P_G,P_H)}{D_1(G,H)}>0.$$

For a converse result, the following is the Hellinger counterpart of Lemma 2.4.4.

Lemma 2.4.12. *Fix $G_0=\sum_{i=1}^{k_0}p_i^0\delta_{\theta_i^0}\in\mathcal{E}_{k_0}(\Theta^\circ)$. Suppose*

a) there exists $\mathcal{N}$ (that possibly depends on G_0) such that $\mu(\mathcal{N})=0$ and for every $x\notin\mathcal{N}$, $f(x|\theta)$ is differentiable w.r.t. θ at $\{\theta_i^0\}_{i=1}^{k_0}$;

b) (2.9a) *with k,θ_i replaced respectively by k_0,θ_i^0 has a nonzero solution $(a_1,b_1,\ldots,a_{k_0},b_{k_0})$;*

c) the density family has common support, i.e. $S=\{x\in\mathfrak{X}|f(x|\theta)>0\}$ does not depend on $\theta\in\Theta$;

d) there exist $\gamma_0>0$ such that $\forall\ 1\le i\le k_0$, $\forall\ 0<\Delta\le\gamma_0$,

$$\left|\frac{f(x|\theta_i^0+a_i\Delta)-f(x|\theta_i^0)}{\Delta\sqrt{f(x|\theta_i^0)}}\right|\le\bar{f}(x),\quad \mu-a.e.\ x\in S\backslash\mathcal{N},$$

where $\bar{f}(x)$ satisfies $\int_{S\backslash\mathcal{N}}\bar{f}^2(x)d\mu<\infty$;

then

$$\lim_{r\to 0}\inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}}\frac{h(P_G,P_H)}{D_1(G,H)}=\liminf_{\substack{G\xrightarrow{W_1}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}}\frac{h(P_G,P_{G_0})}{D_1(G,G_0)}=0 \tag{2.19}$$

Remark 2.4.13. Similar to Remark 2.4.5, one may replace the condition d) in the preceding lemma by the following weaker condition: Condition d'): there exist $\gamma_0>0$ such that $\forall\ 1\le i\le k_0$, $\forall\ 0<\Delta\le\gamma_0$,

$$\left|\frac{f(x|\theta_i^0+a_i\Delta)-f(x|\theta_i^0)}{\Delta\sqrt{f(x|\theta_i^0)}}\right|\le\bar{f}_\Delta(x),\quad \mu-a.e.\ x\in S\backslash\mathcal{N},$$

where $\bar{f}_\Delta(x)$ satisfies $\lim_{\Delta\to 0^+}\int_{S\backslash\mathcal{N}}\bar{f}_\Delta^2(x)d\mu=\int_{S\backslash\mathcal{N}}\lim_{\Delta\to 0^+}\bar{f}_\Delta^2(x)d\mu<\infty$. □

2.4.2 Finer characterizations

In order to verify if the first-order identifiability condition is satisfied for a given probability kernel family $\{f(x|\theta)|\theta\in\Theta\}$, according to Definition 2.4.1 one needs to check that system of equations (2.9a) and (2.9b) does not have non-zero solutions. For many common probability kernel

families, the presence of normalizing constant can make this verification challenging, because the normalizing constant is a function of θ, which has a complicated form or no closed form, and its derivative can also be complicated. Fortunately, the following lemma shows that under a mild condition one only needs to check for the family of kernel $\{f(x|\theta)\}$ defined up to a function of θ that is constant in x. Moreover, under additional mild assumptions, the equation (2.9b) can also be dropped from the verification.

Lemma 2.4.14. *Suppose for every x in the μ-positive subset $\mathfrak{X}\backslash\mathcal{N}$ for some $\mathcal{N} \in \mathcal{A}$, $f(x|\theta)$ is differentiable w.r.t. θ at $\{\theta_i\}_{i=1}^k$. Let $g(\theta)$ be a positive differentiable function on Θ° and define $\tilde{f}(x|\theta) = g(\theta)f(x|\theta)$.*

a) (2.9a) *has only the zero solution if and only if* (2.9a) *with f replaced by $\tilde{f}$ has only the zero solution.*

b) Suppose $\mu(\mathcal{N}) = 0$. For a fixed set $\{a_i\}_{i=1}^k \subset \mathbb{R}^q$ and for each $i \in [k]$ there exists $\gamma(\theta_i, a_i) > 0$ such that for any $0 < \Delta \leq \gamma(\theta_i, a_i)$,

$$\left|\frac{f(x|\theta_i + a_i\Delta) - f(x|\theta_i)}{\Delta}\right| \leq \bar{f}(x|\theta_i, a_i) \quad \mu - a.e.\ \mathfrak{X} \tag{2.20}$$

where $\bar{f}(x|\theta_i, a_i)$ is μ-integrable, then $(a_1, b_1, \ldots, a_k, b_k)$ is a solution of (2.9a) *if and only if it's a solution of the system of equations* (2.9a), (2.9b). *Here $\gamma(\theta_i, a_i)$ and $\bar{f}(x|\theta_i, a_i)$ depend on θ_i and a_i. Moreover,* (2.20) *holds for some μ-integrable $\bar{f}$ if and only if the same inequality with f on the left side replaced by $\tilde{f}$ holds for some μ-integrable $\bar{f}_1$.*

c) Suppose the conditions in b) hold for any set $\{a_i\}_{i=1}^k$. Then (2.9a) *has the same solutions as the system of equations* (2.9a), (2.9b). *Hence, the family $\{f(x|\theta)\}_{\theta\in\Theta}$ is $(\{\theta_i\}_{i=1}^k, \mathcal{N})$ first-order identifiable if and only if* (2.9a) *with f replaced by $\tilde{f}$ has only the zero solution.*

Note similar extension as in Remark 2.4.5 can be made in Lemma 2.4.14 b) and c).

Remark 2.4.15. Part b), or Part c), of Lemma 2.4.14 shows that under some differentiability condition (i.e. $\mu(\mathcal{N}) = 0$) and some regularity condition on the density $f(x|\theta)$ to ensure the exchangibility of limit and the integral, in the definition of $(\{\theta_i\}_{i=1}^k, \mathcal{N})$ first identifiable (2.9b) adds no additional constraint and is redundant. In this case for first-order identifiability, we can simply check whether (2.9a) has only zero solution or not. In addition, according to Part c) of Lemma 2.4.14, for first-order identifiability it is sufficient to check whether (2.9a) with f replaced by $\tilde{f}$ has only zero solution or not, provided that the $\mu(\mathcal{N}) = 0$ for $\mathcal{N}$ corresponds to $\tilde{f}$ and (2.20) with f on the left side replaced by $\tilde{f}$ hold. □

Probability kernels that belong to the exponential families of distribution are frequently employed in practice. For these kernels, there is a remarkable equivalence among the first-order identifiability condition and the inverse bounds for both variational distance and the Hellinger distance.

Lemma 2.4.16. *Suppose that the probability kernel P_θ has a density function f in the full rank exponential family, given in its canonical form $f(x|\theta) = \exp(\langle\theta, T(x)\rangle - A(\theta))h(x)$ with $\theta \in \Theta$, the natural parameter space. Then* (2.9a) *has the same solutions as the system of equations* (2.9a), (2.9b). *Moreover for a fixed $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$ the following five statements are equivalent:*

a) $\lim_{r\to 0} \inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}} \frac{V(P_G,P_H)}{D_1(G,H)} > 0.$

b) $\liminf_{\substack{G\xrightarrow{W_1} G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G,P_{G_0})}{D_1(G,G_0)} > 0;$

c) $\lim_{r\to 0} \inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}} \frac{h(P_G,P_H)}{D_1(G,H)} > 0.$

d) $\liminf_{\substack{G\xrightarrow{W_1} G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{h(P_G,P_{G_0})}{D_1(G,G_0)} > 0;$

e) With k, θ_i replaced respectively by k_0, θ_i^0, equation (2.9a) *has only the zero solution.*

In the last result, the exponential family is in its canonical form. The same conclusions hold for the exponential family represented in general parameterizations.

Lemma 2.4.17. *Consider the probability kernel P_θ has a density function f in the full rank exponential family, $f(x|\theta) = \exp\left(\langle\eta(\theta), T(x)\rangle - B(\theta)\right) h(x)$. Suppose the map $\eta : \Theta \to \eta(\Theta) \subset \mathbb{R}^q$ is a homeomorphism*[1]*. Fix $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$. Suppose the Jacobian matrix of the function $\eta(\theta)$, denoted by $J_\eta(\theta) := (\frac{\partial\eta^{(i)}}{\partial\theta^{(j)}}(\theta))_{ij}$ exists and is full rank at θ_i^0 for $i \in [k_0]$. Then with k, θ_i replaced respectively by k_0, θ_i^0,* (2.9a) *has the same solutions as the system of equations* (2.9a), (2.9b). *Moreover the b), d) and e) as in Lemma 2.4.16 are equivalent. If in addition $J_\eta(\theta)$ exists and is continuous in a neighborhood of θ_i^0 for each $i \in [k_0]$, then the equivalence relationships of all the five statements in Lemma 2.4.16 hold.*

Despite the simplicity of kernels in the exponential families, classical and/or first-order identifiability is not always guaranteed. For instance, it is well-known and can be checked easily that the mixture of Bernoulli distributions is not identifiable in the classical sense. We will discuss the Bernoulli kernel in the context of mixtures of product distributions in Example 2.5.10. The following example is somewhat less well-known.

[1] A homeomorphism is a continuous function between topological spaces that has a continuous inverse function.

Example 2.4.18 (Two-parameter Gamma kernel)**.** Consider the gamma distribution $f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\mathbf{1}_{(0,\infty)}(x)$ with $\theta = (\alpha,\beta) \in \Theta := \{(\alpha,\beta)|\alpha > 0, \beta > 0\}$ and the dominating measure μ is the Lebesgue measure on $\mathfrak{X} = \mathbb{R}$. This is a full rank exponential family. For $k_0 \geq 2$ define $\mathcal{G} \subset \mathcal{E}_{k_0}(\Theta^\circ) = \mathcal{E}_{k_0}(\Theta)$ as

$$\mathcal{G} := \{G \in \mathcal{E}_{k_0}(\Theta)|G = \sum_{i=1}^{k_0} p_i\delta_{\theta_i} \text{ and there exist } i \neq j \text{ such that } \theta_j - \theta_i = (1,0)\}.$$

For any $G_0 = \sum_{i=1}^{k_0} p_i\delta_{\theta_i^0} \in \mathcal{G}$, let $i_0 \neq j_0$ be such that $\theta_{j_0}^0 - \theta_{i_0}^0 = (1,0)$, i.e. $\alpha_{j_0}^0 = \alpha_{i_0}^0 + 1$ and $\beta_{j_0}^0 = \beta_{i_0}^0$. Then observing

$$\frac{\partial}{\partial\beta}f(x|\alpha,\beta) = \frac{\alpha}{\beta}f(x|\alpha,\beta) - \frac{\alpha}{\beta}f(x|\alpha+1,\beta),$$

$(a_1, b_1, \ldots, a_{k_0}, b_{k_0})$ with $a_{i_0} = (0, \beta_{i_0}/\alpha_{i_0})$, $b_{i_0} = -1$, $b_{j_0} = 1$ and the rest to be zero is a nonzero solution of the system of equations (2.9a), (2.9b) with k, θ_i replaced respectively by k_0, θ_i^0. Write gamma distribution in exponential family as in Lemma 2.4.17 with $\eta(\theta) = (\alpha - 1, \beta)$ and $T(x) = (\ln x, -x)$. Since $\eta(\theta)$ satisfies all the conditions in Lemma 2.4.17, hence

$$\liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{h(P_G,P_{G_0})}{D_1(G,G_0)} = \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G,P_{G_0})}{D_1(G,G_0)} = 0.$$

This implies that even if $V(p_G, p_{G_0})$ vanishes at a fast rate, $D_1(G, G_0)$ may not.

Finite mixtures of Gamma were investigated by [HN16a], who called $\mathcal{G}$ is a *pathological* set of parameter values to highlight the effects of *weak identifiability* (more precisely, the violation of first-order identifiability conditions) on the convergence behavior of model parameters when the parameter values fall in $\mathcal{G}$. On the other hand, for $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)\backslash\mathcal{G}$, it is shown in the proof of Proposition 5.1 (a) in [HN16a] that (2.9a) with k, θ_i replaced respectively by k_0, θ_i^0 has only the zero solution.[2] Thus by Lemma 2.4.17,

$$\liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{h(P_G,P_{G_0})}{D_1(G,G_0)} \geq \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G,P_{G_0})}{D_1(G,G_0)} > 0.$$

Notice that a) and c) in Lemma 2.4.17 also holds but is omitted here. Thus, outside of pathological set $\mathcal{G}$ the convergence rate of mixture density p_G towards p_{G_0} is carried over to the convergence of G

[2]Their original proof there only works under the stringent condition $\alpha \geq 1$ for the parameter space. But multiplying their (26) by x should reach the same conclusion for the general case $\alpha > 0$. A direct proof is also straightforward by using Lemma 2.10.4 b) and is similar to Example 2.5.12. In addition, by applying Lemma 2.4.17 we produce additional results on Hellinger distance and drop the unnecessary condition $\alpha \geq 1$ in the parameter space.

toward G_0 under D_1. It is the uncertainty about whether the true mixing measure G_0 is pathological or not that that makes parameter estimation highly inefficient. Given m-i.i.d. from a finite mixture of Gamma distributions, where the number of components k_0 is given, [HN16a] established minimax bound for estimating G that is slower than any polynomial rate m^{-r} for any $r \geq 1$ under W_r metric. $\square$

We end this section with several remarks, motivated by a concern for the situation of weak identifiability. It may be of interest to devise an efficient parameter estimation method (by, perhaps, a clever regularization or reparameterization technique) that may help to overcome the lack of first-order identifiability. We are not aware of a general way to achieve this. Absent of such methods, a promising direction for the statistician to take is to simply collect more data: not only by increasing the number of i.i.d. observation of m, but also by increasing the number of repeated measurements. Finite mixtures of product distributions usually arise in this practical context: when one deals with a highly heterogeneous data population which is made up of many latent subpopulations carrying distinct patterns, it is often possible to collect observations presumably coming from the same subpopulation, even if one is uncertain about the mixture component that a subpopulation may be assigned to. Thus, one may aim to collect m independent sequences of N exchangeable observations, and assume that they are sampled from a finite mixture of N-product distribution denoted by $P_{G,N}$.

One natural question to ask is, how does increasing the number N of repeated measurements (i.e., the length of exchangeable sequences) help to overcome the lack of strong identifiability such as our notion of first-order identifiability. This question can be made precise in light of Lemma 2.4.2: whether there exist a natural number $n_1 \geq 1$ such that the following inverse bound holds for any $N \geq n_1$

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_1(G, G_0)} > 0. \tag{2.21}$$

Observe that since $V(P_{G,N}, P_{G_0,N})$ increases in N while the denominator $D_1(G, G_0)$ is fixed in N, if (2.21) holds for some $N = n_1$, then it also holds for all $N \geq n_1$. Moreover, what can we say about the role of N in parameter estimation in presence of such inverse bounds? In the next section we will address these questions by studying conditions under which inverse bounds involving mixtures of product distributions can be established. Such theory will also be used to derive tight learning rates for mixing measure G from a collection of exchangeable sequences of observations.

2.5 Inverse bounds for mixtures of product distributions

Consider a family of probability distributions $\{P_\theta\}_{\theta\in\Theta}$ on some measurable space $(\mathfrak{X},\mathcal{A})$ where θ is the parameter of the family and $\Theta\subset\mathbb{R}^q$ is the parameter space. This yields the N-product probability kernel on $(\mathfrak{X}^N,\mathcal{A}^N)$, which is denoted by $\{P_{\theta,N}:=\bigotimes^N P_\theta\}_{\theta\in\Theta}$. For any $G=\sum_{i=1}^k p_i\delta_{\theta_i}\in\mathcal{E}_k(\Theta)$ as mixing measure, the resulting finite mixture for the N-product families is a probability measure on $(\mathfrak{X}^N,\mathcal{A}^N)$, namely, $P_{G,N}=\sum_{i=1}^k p_i P_{\theta_i,N}$.

The main results of this section are stated in Theorem 2.5.7 and Theorem 2.5.14. These theorems establish the following inverse bound under certain conditions of probability kernel family $\{P_\theta\}_{\theta\in\Theta}$ and some time that of G_0: for a fixed $G_0\in\mathcal{E}_{k_0}(\Theta^\circ)$ there holds

$$\liminf_{N\to\infty}\ \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}}\ \frac{V(P_{G,N},P_{G_0,N})}{D_N(G,G_0)}>0. \tag{2.22}$$

By contrast, an easy upper bound on the left hand side of (2.22) holds generally (cf. Lemma 2.7.1):

$$\sup_{N\ge1}\ \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}}\ \frac{V(P_{G,N},P_{G_0,N})}{D_N(G,G_0)}\le 1/2. \tag{2.23}$$

In fact, a strong inverse bound can also be established:

$$\liminf_{N\to\infty}\lim_{r\to0}\ \inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}}\ \frac{V(P_{G,N},P_{H,N})}{D_N(G,H)}>0. \tag{2.24}$$

These inverse bounds relate to the positivity of a suitable notion of curvature on the space of mixtures of product distributions, and will be shown to have powerful consequences. It's easy to see that (2.24) implies (2.22).

Section 2.5.2 is devoted to proving these bounds for P_θ belonging to exponential families of distributions. In Section 2.5.3 the inverse bounds are established for very general probability kernel families, where $\mathfrak{X}$ may be an abstract space and no parametric assumption on P_θ will be required. Instead, we appeal to a set of mild regularity conditions on the characteristic function of a push-forward measure produced by a measurable map T acting on the measure space $(\mathfrak{X},\mathcal{A},P_\theta)$. We will see that this general theory enables the study for a very broad range of mixtures of product distributions for exchangeable sequences.

2.5.1 Implications on classical and first-order identifiability

Before presenting the section's main theorems, let us explicate some immediate implications of their conclusions expressed by inequalities (2.22) and (2.24). These inequalities contain detailed information about convergence behavior of de Finetti's mixing measure G toward G_0, an useful application of which will be demonstrated in Section 2.6. For now, we simply highlight striking implications on the basic notions of identifiability of mixtures of distributions investigated in Section 2.4. Note that no assumption on classical or first-order identifiability is required in the statement of the theorems establishing (2.22) or (2.24). In plain terms these inequalities entail that by increasing the number N of exchangeable measurements, the resulting mixture of N-product distributions becomes identifiable in both classical and first-order sense, even if it is not initially so, i.e., when $N = 1$ or small.

To make our statement precise, for a fixed $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$, define

$$
\begin{aligned}
n_0 := n_0(G_0) &:= \min\left\{ n \geq 1 \middle| \forall G \in \cup_{k=1}^{k_0} \mathcal{E}_k(\Theta^\circ) \setminus \{G_0\}, P_{G,n} \neq P_{G_0,n} \right\}, \\
n_1 := n_1(G_0) &:= \min\left\{ n \geq 1 \middle| \liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,n}, P_{G_0,n})}{D_1(G, G_0)} > 0 \right\}, \\
n_2 := n_2(G_0) &:= \min\left\{ n \geq 1 \middle| \lim_{r \to 0} \inf_{\substack{G,H \in B_{W_1}(G_0,r) \\ G \neq H}} \frac{V(P_{G,n}, P_{H,n})}{D_1(G, H)} > 0 \right\}. \qquad (2.25)
\end{aligned}
$$

n_0 is called minimal zero-order identifiable length, or 0-*identifiable length* for short. n_1 is called minimal first-order identifiable length, or 1-*identifiable length* for short. Since $W_1(G, G_0) \asymp_{G_0} D_1(G, G_0)$ in small neighborhood of G_0 (cf. Lemma 2.3.2 d)), the two metrics can be exchangeable in the denominator of the above definition for n_1 and n_2. Note that n_0 describes a global property of classical identifiability, a notion determined mainly by the algebraic structure of the mixture model's kernel family and its parameterization. (This is also known as "strict identifiability" as opposed to "generic identifiability", cf., e.g., [AMR09]). On the other hand, both n_1 and n_2 characterize a local behavior of mixture densities $p_{G,N}$ near a certain $p_{G_0,N}$, a notion that relies primarily on regularity conditions on the kernel, as we shall see in what follows.

The following proposition provides the link between classical identifiability and strong notions of local identifiability provided either (2.22) or (2.24) holds. In a nutshell, as N gets large, the two types of identifiability can be connected by the force of the central limit theorem, which is one of the key ingredients in the proof of the inverse bounds. Define two related and useful quantities: for

any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$

$$\underline{\mathbf{N}}_1 := \underline{\mathbf{N}}_1(G_0) := \min\left\{ n \geq 1 \middle| \inf_{N \geq n} \liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)} > 0 \right\} \tag{2.26}$$

$$\underline{\mathbf{N}}_2 := \underline{\mathbf{N}}_2(G_0) := \min\left\{ n \geq 1 \middle| \inf_{N \geq n} \lim_{r \to 0} \inf_{\substack{G,H \in B_{W_1}(G_0,r) \\ G \neq H}} \frac{V(P_{G,N}, P_{H,N})}{D_N(G, H)} > 0 \right\}. \tag{2.27}$$

Note that (2.22) means $\underline{\mathbf{N}}_1(G_0) < \infty$, while (2.24) means $\underline{\mathbf{N}}_2(G_0) < \infty$. The following proposition collects connections among n_0, n_1, n_2, $\underline{\mathbf{N}}_1$ and $\underline{\mathbf{N}}_2$.

Proposition 2.5.1. *a) Consider any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$, then $n_1(G_0) \leq n_2(G_0)$. Moreover, there exists $r := r(G_0) > 0$ such that*

$$\sup_{G \in B_{W_1}(G_0, r)} n_1(G) \leq n_2(G_0).$$

b) Consider any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$. If $\underline{N}_1(G_0) < \infty$, then $n_1(G_0) = \underline{N}_1(G_0) < \infty$.
If $\underline{N}_2(G_0) < \infty$ then $n_2(G_0) = \underline{N}_2(G_0) < \infty$. In particular, the first or the second conclusion holds if (2.22) *or* (2.24) *holds respectively.*

c) There holds

$$\sup_{G \in \bigcup_{k \leq k_0} \mathcal{E}_k(\Theta^\circ)} n_0(G) \leq \sup_{G \in \mathcal{E}_{2k_0}(\Theta^\circ)} n_1(G).$$

d) Suppose the kernel family P_θ admits density $f(\cdot|\theta)$ with respect to a dominating measure μ on $\mathfrak{X}$. Fix $G_0 = \sum_{i=1}^{k_0} \in \mathcal{E}_{k_0}(\Theta^\circ)$. Moreover, assume all conditions in Corollary 2.4.6 hold for the $f(x|\theta)$. Then, $n_2(G_0) = n_1(G_0)$.

e) Suppose that (2.22) *holds for every $G_0 \in \cup_{k \leq 2k_0} \mathcal{E}_k(\Theta^\circ)$, where Θ is a compact set. Moreover, suppose all conditions in part d) are satisfied for every $G_0 \in \cup_{k \leq 2k_0} \mathcal{E}_k(\Theta)$. Then*

$$\sup_{G \in \cup_{k \leq k_0} \mathcal{E}_k(\Theta^\circ)} n_0(G) \leq \sup_{G \in \mathcal{E}_{2k_0}(\Theta^\circ)} n_1(G) < \infty.$$

f) Suppose that (2.24) *holds for every $G_0 \in \cup_{k \leq 2k_0} \mathcal{E}_k(\Theta^\circ)$, where Θ is a compact set. Then the conclusion of part e) holds.*

Remark 2.5.2. Part a) and part b) of Proposition 2.5.1 highlight an immediate significance of inverse bounds (2.22) and (2.24): they establish pointwise finiteness of 1-identifiable length $n_1(G_0)$. Moreover, under the additional condition on first-order identifiability, one can have the following

strong result as an immediate consequence: Consider any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$. If (2.11) and (2.22) hold, then $n_1(G_0) = \underline{\mathrm{N}}_1(G_0) = 1$. If (2.12) and (2.24) hold, then $n_1(G_0) = \underline{\mathrm{N}}_1(G_0) = n_2(G_0) = \underline{\mathrm{N}}_2(G_0) = 1$. □

Remark 2.5.3. Part e) and part f) in the above proposition establish a rather surprising consequence of inverse bounds (2.22) and (2.24), provided that the domain of support Θ is compact: they yield the finiteness of both 0-identifiable length $n_0(G)$ and 1-identifiable length $n_1(G)$ *uniformly* over subsets of mixing measures with finite number of support points. In particular, as long as (2.22) or (2.24) (along with some regularity conditions in the former) holds for every $G_0 \in \cup_{k\leq 2k_0}\mathcal{E}_k(\Theta^\circ)$, then $P_{G,N}$ will be strictly identifiable and first-identifiable on $\cup_{k\leq k_0}\mathcal{E}_k(\Theta^\circ)$ for sufficiently large N. That is, taking product helps in making the kernel identifiable in a strong sense. As we shall see in the next subsection, (2.24) holds for every $G_0 \in \bigcup_{k=1}^{\infty}\mathcal{E}_k(\Theta^\circ)$ when $\{P_\theta\}$ belongs to full rank exponential families of distributions. This inverse bound also holds for a broad range of probability kernels beyond the exponential families.

Proof of Proposition 2.5.1: a) It is sufficient to assume that $n_2 = n_2(G_0) < \infty$. Then there exists $r_0 > 0$ such that

$$\inf_{\substack{G,H\in B_{W_1}(G_0,r_0)\\ G\neq H}} \frac{V(P_{G,n_2}, P_{H,n_2})}{D_1(G,H)} > 0$$

Then fixing G in the preceding display yields $n_1(G) \leq n_2(G_0)$ and the proof is complete since G is arbitrary in $B_{W_1}(G_0, r_0)$.

b) By the definition of $\underline{\mathrm{N}}_1$,

$$\liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,\underline{\mathrm{N}}_1}, P_{G_0,\underline{\mathrm{N}}_1})}{D_1(G,G_0)} \geq \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,\underline{\mathrm{N}}_1}, P_{G_0,\underline{\mathrm{N}}_1})}{D_{\underline{\mathrm{N}}_1}(G,G_0)} > 0, \tag{2.28}$$

which entails that $n_1 \leq \underline{\mathrm{N}}_1$. On the other hand, for any $N \in [n_1, \underline{\mathrm{N}}_1]$ we have

$$\begin{aligned}\liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G,G_0)} &\geq \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{1}{\sqrt{N}}\frac{V(P_{G,n_1}, P_{G_0,n_1})}{D_1(G,G_0)}\\ &\geq \frac{1}{\sqrt{\underline{\mathrm{N}}_1}} \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,n_1}, P_{G_0,n_1})}{D_1(G,G_0)} > 0,\end{aligned}$$

which entails $\underline{\mathrm{N}}_1 \leq n_1$. Thus $\underline{\mathrm{N}}_1 = n_1$. The proof of $n_2 = \underline{\mathrm{N}}_2 < \infty$ is similar.

c) If suffices to prove for the case $\bar{n} := \sup_{G\in\mathcal{E}_{2k_0}(\Theta^\circ)} n_1(G) < \infty$. Take any $G \in \mathcal{E}_k(\Theta^\circ)$ for $1 \leq k \leq k_0$, suppose that $n_0(G) > \bar{n}$. Then there exists a $G_1 \in \mathcal{E}_{\bar{k}}(\Theta^\circ)$ for some $1 \leq \bar{k} \leq k_0$

such that $P_{G,\bar{n}} = P_{G_1,\bar{n}}$ but $G_1 \neq G$. Collecting the supporting atoms of G and G_1, there are at most $2k_0$ of those, and denote them by $\theta_1^0, \ldots, \theta_{k'}^0 \in \Theta^\circ$. Supplement these with a set of atoms $\{\theta_i^0\}_{i=k'+1}^{2k_0}$ to obtain a set of distinct $2k_0$ atoms denoted by $\{\theta_i^0\}_{i=1}^{2k_0}$. Now take G_0 to be any discrete probability measure supported by $\theta_1^0, \ldots, \theta_{2k_0}^0$. Since $P_{G,\bar{n}} = P_{G_1,\bar{n}}$, the condition of Lemma 2.11.1 for G_0 is satisfied and thus

$$\liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,\bar{n}}, P_{G_0,\bar{n}})}{D_1(G, G_0)} = 0.$$

But this contradicts with the definition of $\bar{n}$.

d) By part a) it suffices to prove for the case $n_1 = n_1(G_0) < \infty$. By Lemma 2.11.3, the product family $\prod_{j=1}^{n_1} f(x_j|\theta)$ satisfies all the conditions in Corollary 2.4.6. Thus by Corollary 2.4.6,

$$\lim_{r \to 0} \inf_{\substack{G,H \in B_{W_1}(G_0,r) \\ G \neq H}} \frac{V(P_{G,n_1}, P_{H,n_1})}{D_1(G, H)} > 0$$

It follows that $n_2(G_0) \leq n_1$, which implies that $n_2(G_0) = n_1(G_0)$ by part a).

e) By part b) and part d), $n_2(G_0) < \infty$ for every $G_0 \in \cup_{k \leq 2k_0} \mathcal{E}_k(\Theta)$. Associated each G_0 with a neighborhood $B_{W_1}(G_0, r(G_0))$ as in part a) such that its conclusion holds. Due to the fact that $\cup_{k \leq 2k_0} \mathcal{E}_k(\Theta)$ is compact and part a), we deduce that $n_1(G)$ is uniformly bounded. Combining with part c) to conclude the proof.

f) By part b) $n_2(G_0) < \infty$ for every $G_0 \in \cup_{k \leq 2k_0} \mathcal{E}_k(\Theta^\circ)$. The rest of the argument is the same as part e).

□

We can further unpack the double infimum limits in its expression of (2.22) to develop results useful for subsequent convergence rate analysis in Section 2.6. First, it is simple to show that the limiting argument for N can be completely removed when N is suitably bounded.

Lemma 2.5.4. *Fix $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$. Suppose* (2.22) *holds. Then for any $N_0 \geq n_1(G_0)$, there exists $c(G_0, N_0)$ depending on G_0 and N_0, such that $\forall G \in \mathcal{E}_{k_0}(\Theta) : W_1(G, G_0) < c(G_0, N_0)$*

$$V(P_{G,N}, P_{G_0,N}) \geq C(G_0) D_N(G, G_0) \quad \forall N \in [n_1(G_0), N_0]$$

where $C(G_0) > 0$ is a constant that depends on G_0.

A key feature of the above claim is that the radius $c(G_0, N_0)$ of the local W_1 ball centered at G_0 over which the inverse bound holds for G depends on N_0, but the multiple constant $C(G_0)$ does not.

Next, given additional conditions, most notably the compactness on the space of mixing measures, we may remove completely the second limiting argument involving G. In other words, we may extend the domain of G on which the inverse bound of the form $V \gtrsim W_1 \gtrsim D_1$ continues to hold, where the multiple constants are suppressed here.

Lemma 2.5.5. *Fix $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$. Consider any Θ_1 a compact subset of Θ containing G_0. Suppose the map $\theta \mapsto P_\theta$ from $(\Theta_1, \|\cdot\|_2)$ to $(\{P_\theta\}_{\theta\in\Theta}, h)$ is continuous. Let $n_1(G_0)$ be given by* (2.25)*. Suppose there exists $n_0 \geq 1$ such that map $G \mapsto P_{G,n_0}$ is identifiable at G_0 on $\bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1))$, i.e. for any $G \in \bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1))\backslash\{G_0\}$, $P_{G,n_0} \neq P_{G_0,n_0}$. If $n_1(G_0) \vee n_0 < \infty$, then*

$$V(P_{G,N}, P_{G_0,N}) \geq C(G_0, \Theta_1) W_1(G, G_0), \quad \forall G \in \bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1),\ \forall N \geq n_1(G_0) \vee n_0,$$

where $C(G_0, \Theta_1) > 0$ is a constant that depends on G_0 and Θ_1.

Finally, a simple and useful fact which allows one to transfer an inverse bound for one kernel family P_θ to another kernel family by means of homeomorphic transformation in the parameter space. If $g(\theta) = \eta$ for some homeomorphic function $g : \Theta \to \Xi \subset \mathbb{R}^q$, for any $G = \sum_{i=1}^k p_i \delta_{\theta_i} \in \mathcal{E}_k(\Theta)$, denote $G^\eta = \sum_{i=1}^k p_i \delta_{g(\theta_i)} \in \mathcal{E}_k(\Xi)$. Given a probability kernel family $\{P_\theta\}_{\theta\in\Theta}$, under the new parameter η define

$$\tilde{P}_\eta = P_{g^{-1}(\eta)}, \quad \forall \eta \in \Xi.$$

Let G^η also denote a generic element in $\mathcal{E}_{k_0}(\Xi)$, and $\tilde{P}_{G^\eta,N}$ be defined similarly as $P_{G,N}$.

Lemma 2.5.6 (Invariance under homeomorphic parameterization with local invertible Jacobian)**.** *Suppose g is a homeomorphism. For $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$, suppose the Jacobian matrix of the function $g(\theta)$, denoted by $J_g(\theta) := (\frac{\partial g^{(i)}}{\partial \theta^{(j)}}(\theta))_{ij}$ exists and is full rank at θ_i^0 for $i \in [k_0]$. Then $\forall N$*

$$\liminf_{\substack{G^\eta \xrightarrow{W_1} G_0^\eta \\ G^\eta \in \mathcal{E}_{k_0}(\Xi)}} \frac{V(\tilde{P}_{G^\eta,N}, \tilde{P}_{G_0^\eta,N})}{D_N(G^\eta, G_0^\eta)} \underset{\gtrsim}{G_0} \liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)}. \tag{2.29}$$

Moreover, if in addition $J_g(\theta)$ exists and is continuous in a neighborhood of θ_i^0 for each $i \in [k_0]$, then $\forall N$

$$\lim_{r\to 0} \inf_{\substack{G^\eta, H^\eta \in B_{W_1}(G_0^\eta, r) \\ G^\eta \neq H^\eta}} \frac{V(\tilde{P}_{G^\eta,N}, \tilde{P}_{H^\eta,N})}{D_1(G^\eta, H^\eta)} \underset{\gtrsim}{G_0} \lim_{r\to 0} \inf_{\substack{G, H \in B_{W_1}(G_0, r) \\ G \neq H}} \frac{V(P_{G,N}, P_{H,N})}{D_1(G, H)}.$$

Lemma 2.5.6 shows that if an inverse bound (2.22) or (2.24) under a particular parametrization is established, then the same inverse bound holds for all other parametrizations that are homeomorphic and that have local invertible Jacobian. This allows one to choose the most convenient

parametrization for a probability kernel family; for instance, one may choose the canonical form for an exponential family or another more convenient parametrization, like the mean parametrization.

2.5.2 Probability kernels in regular exponential family

We now present the first inverse bound for the mixture of products of exponential family distributions. Suppose that $\{P_\theta\}_{\theta\in\Theta}$ is a full rank exponential family of distributions on $\mathfrak{X}$. Adopting the notational convention for canonical parameters of exponential families, we assume P_θ admits a density function with respect to a dominating measure μ, namely $f(x|\theta)$ for $\theta \in \Theta$.

Theorem 2.5.7. *Suppose that the probability kernel $\{f(x|\theta)\}_{\theta\in\Theta}$ is in a full rank exponential family of distributions in canonical form as in Lemma 2.4.16. For any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$,* (2.22) *and* (2.24) *hold.*

In the last theorem the exponential family is in its canonical form. The next corollary shows the same conclusions hold for the exponential family in general form under mild conditions.

Corollary 2.5.8. *Consider the probability kernel P_θ has a density function f in the full rank exponential family, $f(x|\theta) = \exp\left(\langle\eta(\theta), T(x)\rangle - B(\theta)\right) h(x)$, where the map $\eta : \Theta \to \eta(\Theta) \subset \mathbb{R}^q$ is a homeomorphism. Suppose that η is continuously differentiable on Θ° and its the Jacobian is of full rank on Θ°. Then, for any $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$,* (2.22) *and* (2.24) *hold.*

As a consequence of Theorem 2.5.7 (more pertinently, Corollary 2.5.8), Lemma 2.5.1 and Lemma 2.4.17, we immediately obtain the following interesting result for which a direct proof may be challenging.

Corollary 2.5.9. *Let the probability kernel $\{f(x|\theta)\}_{\theta\in\Theta}$ be in a full rank exponential family of distributions as in Corollary 2.5.8 and suppose that all conditions there hold. Then for any $k_0 \geq 1$ and for any $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$, $n_1(G_0) = n_2(G_0) = \underline{N}_1(G_0) = \underline{N}_2(G_0)$ are finite. Moreover,*

$$\sum_{i=1}^{k_0}\left(a_i^T\nabla_\theta\prod_{n=1}^{N} f(x_n|\theta_i^0) + b_i\prod_{n=1}^{N} f(x_n|\theta_i^0)\right) = 0, \quad \bigotimes^{N}\mu - a.e.\ (x_1,\ldots,x_N) \in \mathfrak{X}^N \tag{2.30}$$

has only the zero solution:

$$b_i = 0 \in \mathbb{R} \text{ and } a_i = \mathbf{0} \in \mathbb{R}^q, \quad \forall 1 \leq i \leq k_0$$

if and only if $N \geq n_1(G_0)$.

Corollary 2.5.9 establishes that for full rank exponential families of distribution specified in Corollary 2.5.8 with full rank Jacobian of $\eta(\theta)$, there is a finite phase transition behavior specified by $n_1(G_0)$ of the N-product in (2.30): the system of equations (2.30) has nonzero solution when $N < n_1(G_0)$ and as soon as $N \geq n_1(G_0)$, it has only the zero solution. This also gives another characterization of $n_1(G_0)$ defined in (2.25) for such exponential families, which also provides a way to compute $n_1(G_0) = \underline{\mathrm{N}}_1(G_0) = n_2(G_0) = \underline{\mathrm{N}}_2(G_0)$. A byproduct is that $n_0(G_0)$ does not depend on the p_i^0 of G_0 since (2.30) only depends on θ_i^0.

We next show two nontrivial examples of mixture models that are either non-identifiable or weakly identifiable, i.e., when $N = 1$, but become first-identifiable by taking products. We work out the details on calculating $n_0(G_0)$ and $n_1(G_0)$ for each of the examples and they should serve as convincing examples to the discussion at the end of Section 2.4.2.

Example 2.5.10 (Bernoulli kernel)**.** Consider the Bernoulli distribution $f(x|\theta) = \theta^x(1-\theta)^{1-x}$ with parameter space $\Theta = (0, 1)$. Here the family is defined on $\mathfrak{X} = \mathbb{R}$ and the dominating measure is $\mu = \delta_0 + \delta_1$. It can be written in exponential form as in Lemma 2.4.17 with $\eta(\theta) = \ln\theta - \ln(1-\theta)$ and $T(x) = x$. It's easy to check that $\eta'(\theta) = \frac{1}{\theta(1-\theta)} > 0$ and thus all conditions in Lemma 2.4.17, Corollary 2.5.8 and Corollary 2.5.9 are satisfied. Thus any of those three results can be applied. In particular we may use the characterization of $n_1(G_0)$ in Corollary 2.5.9 to compute $n_1(G_0)$.

For the n-fold product, the density $f_n(x_1, x_2, \ldots, x_n|\theta) := \prod_{j=1}^n f(x_j|\theta) = \theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j}$. Then the derivative w.r.t. θ of $f_n(x_1, x_2, \ldots, x_n|\theta)$ is

$$\begin{aligned}&\frac{\partial}{\partial\theta} f_n(x_1, \ldots, x_n|\theta)\\ &= \left(\sum_{j=1}^n x_j\right)\theta^{\sum_{j=1}^n x_j - 1}(1-\theta)^{n-\sum_{j=1}^n x_j} - \left(n - \sum_{j=1}^n x_j\right)\theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j - 1}.\end{aligned}$$

We now compute $n_1(G)$ for any $G = \sum_{i=1}^k p_i\delta_{\theta_i} \in \mathcal{E}_k(\Theta)$. Notice the support of f is $\{0, 1\}$ and hence the support of f_n is $\{0, 1\}^n$. Thus (2.30) with N, θ_i^0 replace respectively by n and θ_i become a system of $n + 1$ linear equations: $\forall j = 0, 1, \ldots, n$

$$\sum_{i=1}^k a_i\left(j(\theta_i)^{j-1}(1-\theta_i)^{n-j} - (n-j)(\theta_i)^j(1-\theta_i)^{n-j-1}\right) + \sum_{i=1}^k b_i(\theta_i)^j(1-\theta_i)^{n-j} = 0. \quad (2.31)$$

As a system of $n + 1$ linear equations with $2k$ unknown variables, it has nonzero solutions when $n + 1 < 2k$. Thus $n_1(G) \geq 2k - 1$.

Let us now verify that $n_1(G) = 2k - 1$ for any $G \in \mathcal{E}_k(\Theta)$. Indeed, for any $G = \sum_{i=1}^k p_i\delta_{\theta_i} \in \mathcal{E}_k(\Theta)$, the system of linear equations (2.31) with $n = 2k-1$ is $A^T z = 0$ with $z = (b_1, a_1, \ldots, b_k, a_k)^T$

and

$$A_{ij} = \begin{cases} f_j(\theta_m) & i = 2m-1 \\ f_j'(\theta_m) & i = 2m \end{cases} \text{ for } j \in [2k], m \in [k],$$

where $f_j(\theta) = \theta^{j-1}(1-\theta)^{n-(j-1)}$ with $n = 2k-1$. By Lemma 2.5.11 d) $\det(A) = \prod_{1\le\alpha<\beta\le k}(\theta_\alpha - \theta_\beta)^4$, with the convention 1 when $k = 1$. Thus, A is invertible and the system of linear equations (2.31) with $n = 2k-1$ has only zero solution. Thus by Corollary 2.5.9 $n_1(G) \le 2k-1$. By the conclusion from last paragraph $n_1(G) = 2k-1$ for any $G \in \mathcal{E}_k(\Theta)$.

We now turn our attention to n_0. For any $G \in \mathcal{E}_k(\Theta)$, there are $2k-1$ parameters to determine it. $f_n(x_1, \ldots, x_n)$ has effective n equations for different value $(x_1, \ldots, x_n)$ since $\sum_{j=1}^n x_j$ can takes $n+1$ values and f_n is a probability density. Thus to have $P_{G,n}$ strictly identifiable for $G \in \mathcal{E}_k(\Theta)$, a necessary condition is that $n \ge 2k-1$ for almost all G under Lebesgue. In fact, in Lemma 2.11.4 part e) it is established that $n_0(G) \ge 2k-1$ for *all* $G \in \mathcal{E}_k(\Theta)$.

Let us now verify that $n_0(G) = 2k-1$ for any $G \in \mathcal{E}_k(\Theta)$. In the following $n = 2k-1$. For any $G = \sum_{i=1}^k p_i\delta_{\theta_i}$ and consider $G' = \sum_{i=1}^k p_i'\delta_{\theta_i'} \in \bigcup_{i=1}^k \mathcal{E}_k(\Theta)$ such that $p_{G',n} = p_{G,n}$. Notice that $G' \in \bigcup_{i=1}^k \mathcal{E}_k(\Theta)$ means that it is possible some of p_i' is zero. $p_{G',n} = p_{G,n}$ implies

$$\sum_{i=1}^k p_i'(\theta_i')^j(1-\theta_i')^{n-j} - \sum_{i=1}^k p_i(\theta_i)^j(1-\theta_i)^{n-j} = 0 \quad \forall j = 0, 1, \cdots, n. \tag{2.32}$$

Note that by multiplying each equation j by $\binom{n}{j}$ and sum them up, we obtain $\sum_{i=1}^k p_i' = \sum_{i=1}^k p_i$. Thus in the above system of equations the equation with $j = n$ (or arbitrary j) can be replaced by $\sum_{i=1}^k p_i' = \sum_{i=1}^k p_i$.

We now show that the only solution is $G' = G$, beginning with the following simple observation. Notice that for a set $\{\xi_i\}_{i=1}^{2k}$ of $2k$ distinct elements in $(0,1)$, the system of linear equations of $y = (y_1, \ldots, y_{k'})$ with $k' \le 2k$:

$$\sum_{i=1}^{k'} y_i(\xi_i)^j(1-\xi_i)^{n-j} = 0 \quad \forall j = 0, 1, \ldots, n = 2k-1$$

has only the zero solution since by setting $\tilde{y}_i = (1-\xi_i)^n y_i$ the system of equations of $\tilde{y}$:

$$\sum_{i=1}^{k'} \tilde{y}_i \left(\frac{\xi_i}{1-\xi_i}\right)^j = 0 \quad \forall j = 0, 1, \ldots, n$$

has its coefficients of the first k' equations forming a non-singular Vandermonde matrix.

If some θ_i is not in $\{\theta_i'\}_{i=1}^k$, then by the observation in last paragraph $p_i = 0$ which contradict with $G \in \mathcal{E}_k(\Theta)$. As a result, $\{\theta_i'\}_{i=1}^k = \{\theta_i\}_{i=1}^k$. Suppose $\theta_{l_i}' = \theta_i$ for $i \in [k]$. Then the system of

equations (2.32) become

$$\sum_{i=1}^{k}(p'_{l_i}-p_i)(\theta_i)^j(1-\theta_i)^{n-j}=0 \quad \forall j=0,1,\ldots,n.$$

Applying the observation from last paragraph again yields $p'_{l_i}=p_i$ for $i\in[k]$. That is, the only solution of (2.32) is $G'=G$. Thus $n_0(G)\leq 2k-1$, which together with the fact that $n_0(G)\geq 2k-1$ yield $n_0(G)=2k-1$ for any $G\in\mathcal{E}_k(\Theta)$.

□

Part d) of the following lemma is used in the previous example on Bernoulli kernel.

Lemma 2.5.11. *a) Let $f(x)$ be a polynomial with f' its derivative. Define $q^{(1)}(x,y)=\frac{f(x)-f(y)}{x-y}$, $q^{(2)}(x,y)=\frac{f'(x)-f'(y)}{x-y}$, $\bar{q}^{(2)}(x,y)=\frac{q^{(1)}(x,y)-f'(y)}{x-y}$, and $\bar{q}^{(3)}(x,y)=\frac{\bar{q}^{(2)}(x,y)-\frac{1}{2}q^{(2)}(x,y)}{x-y}$. Then $q^{(1)}(x,y)$, $q^{(2)}(x,y)$, $\bar{q}^{(2)}(x,y)$ and $\bar{q}^{(3)}(x,y)$ are all multivariate polynomials.*

b) Let $f_j(x)$ be a polynomial and $f'_j(x)$ its derivative for $j\in[2k]$ for a positive integer k. For $x_1,\ldots,x_k\in\mathbb{R}$ define $A^{(k)}(x_1,\ldots,x_k)\in\mathbb{R}^{(2k)\times(2k)}$ by

$$A^{(k)}_{ij}(x_1,\ldots,x_k)=\begin{cases} f_j(x_m) & i=2m-1 \\ f'_j(x_m) & i=2m \end{cases} \text{ for } j\in[2k], m\in[k].$$

Then for any $k\geq 2$, $det(A^{(k)}(x_1,\ldots,x_k))=g_k(x_1,\ldots,x_k)\prod_{1\leq\alpha<\beta\leq k}(x_\alpha-x_\beta)^4$, where g_k is some multivariate polynomial.

c) For the special case $f_j(x)=f_j(x|k)=x^{j-1}$, $A^{(k)}(x_1,\ldots,x_k)$ defined in part b) has determinant $det(A^{(k)}(x_1,\ldots,x_k))=\prod_{1\leq\alpha<\beta\leq k}(x_\alpha-x_\beta)^4$, with the convention 1 when $k=1$.

d) For the special case $f_j(x)=f_j(x|k)=x^{j-1}(1-x)^{n-(j-1)}$ with $n=2k-1$, $A^{(k)}(x_1,\ldots,x_k)$ defined in part b) has determinant $det(A^{(k)}(x_1,\ldots,x_k))=\prod_{1\leq\alpha<\beta\leq k}(x_\alpha-x_\beta)^4$, with the convention 1 when $k=1$.

Example 2.5.12 (Continuation on two-parameter Gamma kernel)**.** Consider the gamma distribution $f(x|\alpha,\beta)$ discussed in Example 2.4.18. Let $k_0\geq 2$ and by Example 2.4.18 for any $G_0\in\mathcal{E}_{k_0}(\Theta)\backslash\mathcal{G}$, $n_1(G_0)=1$ and for any $G_0\in\mathcal{G}$, where we recall that $\mathcal{G}$ denotes the pathological subset of the Gamma mixture's parameter space,

$$\liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}}\frac{h(P_G,P_{G_0})}{D_1(G,G_0)}=\liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}}\frac{V(P_G,P_{G_0})}{D_1(G,G_0)}=0.$$

This means $n_1(G_0) \geq 2$ for $G_0 \in \mathcal{G}$.

We now show that for $G_0 \in \mathcal{E}_{k_0}(\Theta)$ $n_1(G_0) \leq 2$ and hence $n_1(G_0) = 2$ for $G_0 \in \mathcal{G}$. Let

$$f_2(x_1, x_2|\alpha, \beta) := f(x_1|\alpha, \beta) f(x_2|\alpha, \beta) = \frac{\beta^{2\alpha}}{(\Gamma(\alpha))^2} (x_1 x_2)^{\alpha-1} e^{-\beta(x_1+x_2)} \mathbf{1}_{(0,\infty)\times(0,\infty)}(x_1, x_2)$$

be the density of the 2-fold product w.r.t. Lebesgue measure on $\mathbb{R}^2$. Let $g(\alpha, \beta) = (\Gamma(\alpha))^2/\beta^{2\alpha}$, which is a differentiable function on Θ and let $\tilde{f}_2(x_1, x_2|\alpha, \beta) := g(\alpha, \beta) f_2(x_1, x_2|\alpha, \beta)$ to be the density without normalization constant. Note that $\frac{\partial}{\partial\alpha}\tilde{f}_2(x_1, x_2|\alpha, \beta) = \tilde{f}_2(x_1, x_2|\alpha, \beta) \ln(x_1 x_2)$ and $\frac{\partial}{\partial\beta}\tilde{f}_2(x_1, x_2|\alpha, \beta) = -(x_1 + x_2)\tilde{f}_2(x_1, x_2|\alpha, \beta)$. Then (2.9a) with f replaced by $\tilde{f}_2$ is

$$\sum_{i=1}^{k_0} \left(a_i^{(\alpha)} \ln(x_1 x_2) - a_i^{(\beta)}(x_1 + x_2) + b_i \right) (x_1 x_2)^{\alpha_i - 1} e^{-\beta_i(x_1+x_2)} = 0. \tag{2.33}$$

Let $\bigcup_{i=1}^{k} \{\beta_i\} = \{\beta'_1, \beta'_2, \cdots, \beta'_k\}$ with $\beta'_1 < \beta'_2 < \ldots < \beta'_{k'}$ where k' is the number of distinct elements. Define $I(\beta') = \{i \in [k] | \beta_i = \beta'\}$. Then (2.33) become for μ-a.e. $x_1, x_2 \in (0, \infty)$

$$\begin{aligned}
0 =& \sum_{j=1}^{k'} \left(\sum_{i \in I(\beta'_j)} \left(a_i^{(\alpha)} \ln(x_1 x_2) - a_i^{(\beta)}(x_1 + x_2) + b_i \right) (x_1 x_2)^{\alpha_i - 1} \right) e^{-\beta'_j(x_1+x_2)} \\
=& \sum_{j=1}^{k'} \left(\sum_{i \in I(\beta'_j)} a_i^{(\alpha)} (x_1 x_2)^{\alpha_i - 1} \ln(x_1) + \right. \\
& \left. \sum_{i \in I(\beta'_j)} \left(a_i^{(\alpha)} \ln(x_2) - a_i^{(\beta)}(x_1 + x_2) + b_i \right) (x_1 x_2)^{\alpha_i - 1} \right) e^{-\beta'_j x_2} e^{-\beta'_j x_1}
\end{aligned}$$

When fixing any x_2 such that in the μ-a.e. set such that the preceding equation holds, by Lemma 2.10.4 b) for any $j \in [k']$, $\sum_{i \in I(\beta'_j)} a_i^{(\alpha)} (x_1 x_2)^{\alpha_i - 1} \equiv 0$ for any $x_1 \neq 0$. Since α_i are distinct for $i \in I(\beta'_j)$ and $x_2 > 0$, $a_i^{(\alpha)} = 0$ for any $i \in I(\beta'_j)$ for any $j \in [k']$. That is $a_i^{(\alpha)} = 0$ for any $i \in [k]$. Analogously fixing x_1 produces $a_i^{(\beta)} = 0$ for any $i \in [k]$. Plug these back into the preceding display and one obtains for μ-a.e. $x_1, x_2 \in (0, \infty)$

$$0 = \sum_{j=1}^{k'} \left(\sum_{i \in I(\beta'_j)} b_i (x_1 x_2)^{\alpha_i - 1} \right) e^{-\beta'_j x_2} e^{-\beta'_j x_1}$$

Fixing any x_2 such that in the μ-a.e. set such that the preceding equation holds, and apply Lemma 2.10.4 b) again to obtain $b_i = 0$ for $i \in [k]$. Thus (2.33) for any $G \in \mathcal{E}_k(\Theta)$ has only the zero

solution. By Lemma 2.4.17, for $G_0 \in \mathcal{E}_{k_0}(\Theta)$

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,2}, P_{G_0,2})}{D_1(G, G_0)} > 0.$$

Thus $n_1(G_0) \leq 2$, and hence $n_1(G_0) = 2$ for any $G_0 \in \mathcal{G}$.

Following an analogous analysis, one can show that $\{f(x|\theta_i)\}_{i=1}^k$ are linear independent for any distinct $\theta_1, \ldots, \theta_k \in \Theta$ for any k, i.e. the equations of $(b_1, \ldots, b_k)$

$$\sum_{i=1}^{k} b_i f(x|\theta_i) = 0 \quad \forall x \in (0, \infty)$$

has only the zero solution. This linear independence immediately imply that p_G is identifiable on $\bigcup_{j=1}^{\infty} \mathcal{E}_j(\Theta)$, i.e. for any $G \in \mathcal{E}_k(\Theta)$ and any $G' \in \mathcal{E}_{k'}(\Theta)$, $p_G \neq p_{G'}$. Thus, $n_0(G) = 1$ for any $G \in \bigcup_{j=1}^{\infty} \mathcal{E}_j(\Theta)$. □

The above examples demonstrate the remarkable benefits of having repeated (exchangeable) measurements: via the N-fold product kernel $\prod_{j=1}^{N} f(x_j|\theta)$ for sufficiently large N, one can completely erase the effect of parameter non-identifiablity in Bernoulli mixtures, and the effect of weak-identifiability in the pathological subset of the parameter spaces in two-parameter Gamma mixtures. We have also seen that it is challenging to determine the 0- or 1-identifiable lengths even for these simple examples of kernels. It is even more so, when we move to a broader class of probability kernels well beyond the exponential families.

2.5.3 General probability kernels

Unlike Section 2.5.2, which specializes to the probability kernels that are in the exponential families, in this section no such parametric assumption will be required. In fact, we shall *not* require that the family of probability distributions $\{P_\theta\}_{\theta \in \Theta}$ on $\mathfrak{X}$ admit a density function. Since the primary object of inference is parameter $\theta \in \Theta \subset \mathbb{R}^q$, the assumptions on the kernel P_θ will center on the existence of a measurable map $T : (\mathfrak{X}, \mathcal{A}) \to (\mathbb{R}^s, \mathcal{B}(\mathbb{R}^s))$ for some $s \geq q$, and regularity conditions on the push-forward measure on $\mathbb{R}^s$: $P_\theta \# T := P_\theta \circ T^{-1}$.

Definition 2.5.13 (Admissible transform)**.** A Borel measurable map/transform $T : \mathfrak{X} \to \mathbb{R}^s$ with $s \geq q$ is admissible with respect to a set $\Theta_1 \subset \Theta^\circ$ if for each $\theta_0 \in \Theta_1$ there exists $\gamma > 0$ and $r \geq 1$ such that T satisfies the following three properties.

(A1) (Moment condition) For $\theta \in B(\theta_0, \gamma) \subset \Theta^\circ$, the open ball centered at θ_0 with radius γ, suppose $\lambda(\theta) = \lambda_\theta = \mathbb{E}_\theta T X_1$ and $\Lambda_\theta := \mathbb{E}_\theta (TX_1 - \mathbb{E}_\theta TX_1)(TX_1 - \mathbb{E}_\theta TX_1)^T$ exist where $X_1 \sim P_\theta$. Moreover, Λ_θ is positive definite on $B(\theta_0, \gamma)$ and is continuous at θ_0.

(A2) (Exchangeability of partial derivatives of characteristic functions) Denote by $\phi_T(\zeta|\theta)$ the characteristic function of the pushforward probability measure $P_\theta \# T$ on $\mathbb{R}^s$, i.e., $\phi_T(\zeta|\theta) := \mathbb{E}_\theta e^{i\langle \zeta, TX_1\rangle}$, where $X_1 \sim P_\theta$. $\frac{\partial \phi_T(\zeta|\theta)}{\partial \theta^{(i)}}$ exists in $B(\theta_0, \gamma)$ and as a function of ζ it is twice continuously differentiable on $\mathbb{R}^s$ with derivatives satisfying: $\forall \theta \in B(\theta_0, \gamma)$

$$\frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \zeta^{(j)} \partial \theta^{(i)}} = \frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \theta^{(i)} \partial \zeta^{(j)}}, \quad \frac{\partial^3 \phi_T(\zeta|\theta)}{\partial \zeta^{(\ell)} \partial \zeta^{(j)} \partial \theta^{(i)}} = \frac{\partial^3 \phi_T(\zeta|\theta)}{\partial \theta^{(i)} \partial \zeta^{(\ell)} \partial \zeta^{(j)}}, \quad \forall \zeta \in \mathbb{R}^s, \forall j, \ell \in [d], \forall i \in [k_0]$$

where the right hand side of both equations exist.

(A3) (Continuity and integrability conditions of characteristic function) $\phi_T(\zeta|\theta)$ as a function of θ is twice continuously differentiable in $B(\theta_0, \gamma)$. There exists $U_1(\theta_0), U_2(\theta_0) < \infty$ such that: for any $i \in [q], j \in [s]$,

$$\sup_{\theta \in B(\theta_0, \gamma)} \max \left\{ \sup_{\zeta \in \mathbb{R}^s} \left| \frac{\partial \phi_T(\zeta|\theta)}{\partial \theta^{(i)}} \right|, \sup_{\|\zeta\|_2 < 1} \left| \frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \zeta^{(j)} \partial \theta^{(i)}} \right| \right\} \leq U_1(\theta_0) < \infty,$$

and for any $i, j \in [q]$,

$$\sup_{\theta \in B(\theta_0, \gamma)} \max \left\{ \int_{\mathbb{R}^s} |\phi_T(\zeta|\theta)|^r \left(1 + \left| \frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \theta^{(j)} \partial \theta^{(i)}} \right| \right) d\zeta, \sup_{\|\zeta\|_2 < 1} \left| \frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \theta^{(j)} \partial \theta^{(i)}} \right| \right\} \leq U_2(\theta_0) < \infty.$$

Theorem 2.5.14. *Fix $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$. Assume that for each θ_i^0, there exists measurable transform $T_i : (\mathfrak{X}, \mathcal{A}) \to (\mathbb{R}^{s_i}, \mathcal{B}(\mathbb{R}^{s_i}))$ that is admissible with respect to $\{\theta_i^0\}_{i=1}^k$ with $s_i \geq q$ such that 1) the mean map $\lambda_i(\theta)$ of T_i defined in (A1) is identifiable at θ_i^0 over the set $\{\theta_i^0\}_{i=1}^{k_0}$, i.e., $\lambda_i(\theta_j^0) \neq \lambda_i(\theta_i^0)$ for any $j \in [k_0] \backslash \{i\}$ and 2) the Jacobian matrix of λ_i is of full column rank at θ_i^0. Then* (2.22) *and* (2.24) *hold.*

The following corollary is a special case of Theorem 2.5.14 when the admissbile transform T_i are the same for all $i = 1, \ldots, k_0$.

Corollary 2.5.15. *Fix $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$. If there exists one measurable transform $T : (\mathfrak{X}, \mathcal{A}) \to (\mathbb{R}^s, \mathcal{B}(\mathbb{R}^s))$ that is admissible with respect to $\{\theta_i^0\}_{i=1}^k$ with $s \geq q$ such that 1) the mean map $\lambda(\theta)$ of T defined in (A1) is identifiable over the set $\{\theta_i^0\}_{i=1}^{k_0}$, i.e., $\lambda(\theta_j^0) \neq \lambda(\theta_i^0)$ for any distinct $i, j \in [k_0]$ and 2) the Jacobian matrix of λ is of full column rank at θ_i^0 for any $i \in [k_0]$. Then* (2.22) *and* (2.24) *hold.*

The proofs of Theorem 2.5.7 and Theorem 2.5.14 occupy the bulk of the chapter. They contain a number of potentially useful techniques. The presentation of these proofs are deferred to Section 2.12. We make the following remarks.

Remark 2.5.16. In many cases one may construct a single measurable transform T that is admissible and is mean identifiable, then Corollary 2.5.15 is useful. Indeed, as will be illustrated by examples, for some simple cases, the identity map T suffices, and for some one dimensional probability families, the moment map $Tx = (x^1, \cdots, x^s)$ suffices. However, for some complicated probability kernel families, constructing a single admissible transform T_i to have their Jacobi matrix of its mean map of full column rank at every atoms θ_i^0 or to have its mean map identifiable at atoms may be challenging. In such cases, Theorem 2.5.14 offers a tool since we may construct a sequence of admissible maps, each of which only needs to have these properties specified at one atom. Although in such cases one might still be able to construct one single admissible map by combining T_i, yet this single map is likely of higher dimension and may be difficult to construct directly without constructing $\{T_i\}_{i=1}^{k_0}$ first. □

Remark 2.5.17. Although Theorem 2.5.14 provides inverse bounds for a very broad range of probability kernels, it seems not straightforward to apply it to nondegenerate discrete distributions on lattice points, like Poisson, Bernoulli, geometric distributions etc. The reason is that for nondegenerate discrete distributions on lattice points, its characteristic function is periodic (cf. Lemma 4 in Chapter XV, Section 1 of [Fel08]), which prevents its characteristic function from being in L^r. That is, it does not satisfy (A3) in the definition of admissible transform. Thus to apply Theorem 2.5.14 to such distributions one has to come up with a measurable transform T which induce distributions over a countable support that is not lattice points. On the contrary, Theorem 2.5.7 can be readily applied to discrete distributions that are in the exponential family, including Poisson, Bernoulli, geometric distributions, etc. □

2.5.4 Examples of non-standard probability kernels

The power of Theorem 2.5.14 lies in its applicability to classes of probability kernels that do not belong to the exponential family of distributions.

Example 2.5.18 (Continuation on uniform probability kernel)**.** In Example 2.4.8 this example has been shown to satisfies inverse bound (2.11) and (2.12) for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$. In the following it will be shown that the uniform distribution family satisfies (2.22) and (2.24) for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$. Note this family is not an exponential family and thus Theorem 2.5.7 or Corollary 2.5.8 is not applicable.

Take the T in Corollary 2.5.15 to be the identity map. Then $\lambda(\theta) = \frac{\theta}{2}$, $\Lambda_\theta = \frac{\theta^2}{12}$. So condition (A1) is satisfied. The characteristic function is

$$\phi(\zeta|\theta) = \frac{e^{\boldsymbol{i}\zeta\theta} - 1}{\boldsymbol{i}\zeta\theta}\mathbf{1}(\zeta \neq 0) + \mathbf{1}(\zeta = 0).$$

One can then calculate

$$\begin{aligned}\frac{\partial}{\partial\theta}\phi(\zeta|\theta) =& \frac{e^{\boldsymbol{i}\zeta\theta}(e^{-\boldsymbol{i}\zeta\theta}-1-(-\boldsymbol{i}\zeta\theta))}{\boldsymbol{i}\zeta\theta^2}\mathbf{1}(\zeta\neq 0),\\ \frac{\partial^2}{\partial\zeta\partial\theta}\phi(\zeta|\theta) =& \frac{-e^{\boldsymbol{i}\zeta\theta}(e^{-\boldsymbol{i}\zeta\theta}-1-(-\boldsymbol{i}\zeta\theta)-(-\boldsymbol{i}\zeta\theta)^2)}{\boldsymbol{i}\zeta^2\theta^2}\mathbf{1}(\zeta\neq 0)+\frac{\boldsymbol{i}}{2}\mathbf{1}(\zeta=0),\\ \frac{\partial^2}{\partial\theta^2}\phi(\zeta|\theta) =& \frac{-2e^{\boldsymbol{i}\zeta\theta}(e^{-\boldsymbol{i}\zeta\theta}-1-(-\boldsymbol{i}\zeta\theta)-\frac{1}{2}(-\boldsymbol{i}\zeta\theta)^2)}{\boldsymbol{i}\zeta\theta^3}\mathbf{1}(\zeta\neq 0),\end{aligned}$$

and verify the condition (A2). To verify (A3) the following inequality (see (9.5) in [Res14])

$$\left|e^{\boldsymbol{i}x}-\sum_{k=0}^{j}\frac{(\boldsymbol{i}x)^k}{k!}\right|\leq 2\frac{|x|^j}{j!}$$

comes handy. It follows that

$$\left|\frac{\partial}{\partial\theta}\phi(\zeta|\theta)\right|\leq\frac{2}{\theta},\quad \left|\frac{\partial^2}{\partial\zeta\partial\theta}\phi(\zeta|\theta)\right|\leq\frac{3}{2},\quad \left|\frac{\partial^2}{\partial\theta^2}\phi(\zeta|\theta)\right|\leq\frac{2|\zeta|}{\theta}.$$

Then $U_1(\theta_0)=2(\frac{2}{\theta_0}+\frac{3}{2})$ suffices. Finally take $r=3$, observe

$$|\phi(\zeta|\theta)|^3\left(1+\left|\frac{\partial^2}{\partial\theta^2}\phi(\zeta|\theta)\right|\right)\leq\begin{cases}1+\frac{2}{\theta} & |\zeta|\leq 1\\ \frac{8}{|\zeta|^3\theta^3}\left(1+\frac{2|\zeta|}{\theta}\right) & |\zeta|>1\end{cases},$$

and one may choose appropriate $U_2(\theta_0)$ such that (A3) holds. We have then verified that the identity map T is admissible on Θ.

It's easy to see that $\lambda(\theta)=\theta/2$ is injective on Θ and that its Jacobian $J_\lambda(\theta)=\frac{1}{2}$ is full rank. Then by Corollary 2.5.15 (2.22) and (2.24) hold for any $G_0\in\mathcal{E}_{k_0}(\Theta)$ for any $k_0\geq 1$. Moreover by Remark 2.5.2, $n_1(G_0)=\underline{\mathrm{N}}_1(G_0)=n_2(G_0)=\underline{\mathrm{N}}_2(G_0)=1$ for any $G_0\in\mathcal{E}_{k_0}(\Theta)$ for any $k_0\geq 1$. □

Example 2.5.19 (Continuation on location-scale exponential distribution kernel)**.** In Example 2.4.10 this example has been shown to satisfy inverse bound (2.11) for any $G_0\in\mathcal{E}_{k_0}(\Theta)$. In the following it will be shown that this family satisfies (2.22) and (2.24) for any $G_0\in\mathcal{E}_{k_0}(\Theta)$. Note this family is not exponential family and thus Theorem 2.5.7 or Corollary 2.5.8 is not applicable.

Take the T in Corollary 2.5.15 to be $Tx=(x,x^2)^T$ as a map from $\mathbb{R}\to\mathbb{R}^2$. Then one may check $\lambda(\xi,\sigma)=(\xi+\sigma,\sigma^2+(\sigma+\xi)^2)^T$. From here one can easily check $\lambda:\Theta\to\mathbb{R}^2$ is injective, its Jacobi determinant $\det(J_\lambda)=2\sigma>0$, which implies J_λ is of full rank on Θ. The characteristic function $\phi_T(\zeta|\xi,\sigma)=e^{\boldsymbol{i}\zeta\xi}/(1-\boldsymbol{i}\sigma\zeta)$. The rest of verification that T is admissible are simple calculations and are omitted. Then by Corollary 2.5.15 (2.22) and (2.24) holds for any $G_0\in\mathcal{E}_{k_0}(\Theta)$

for any $k_0 \geq 1$. Moreover by Remark 2.5.2, $n_1(G_0) = \underline{\mathrm{N}}_1(G_0) = 1$ for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$ for any $k_0 \geq 1$. □

Example 2.5.20 (P_θ is itself a mixture distribution)**.** Here, we consider the situation where P_θ is a rather complex object: it is itself a mixture distribution. With this example we are moving from a standard mixture of product distributions to hierarchical models (which describe mixtures of mixture distributions). Such models are central tools in Bayesian statistics. Theorem 2.5.7 or Corollary 2.5.8 is obviously not applicable in this example, which indeed requires the full strength of Theorem 2.5.14 or Corollary 2.5.15. The application of the theorem, however, is non-trivial requiring the development of tools for evaluating oscillatory integrals of interest. Such technical tools also prove useful in other contexts (such as Example 2.5.21). Due to the technical nature we defer a complete treatment of the current example to Section 2.8. □

Example 2.5.21 (P_θ is a mixture of Dirichlet processes)**.** This example illustrates the applicability of our theory to models using probability kernels defined in abstract spaces. Such kernels are commonly found in nonparametric Bayesian modeling literature [HHMW10, GvdV17]. In particular, in our specification of mixture of product distributions we will employ Dirichlet processes as the basic building block [Fer73, Ant74]. We defer a treatment of this example to Section 2.8. □

2.6 Posterior contraction of de Finetti's mixing measures

2.6.1 Data are equal-length exchangeable sequences

Given m independent sequences of exchangeable observations of equal length N, $X^i_{[N]} = (X_{i1}, \cdots, X_{iN}) \in \mathfrak{X}^N$ for $i = 1, 2, \cdots, m$. Each sequence $X^i_{[N]}$ is assumed to be a sample drawn from a mixture of $N-$product distributions $P_{G,N}$ for some "true" mixing measure $G = G_0 \in \mathcal{E}_{k_0}(\Theta)$. A Bayesian statistician endows upon $(\mathcal{E}_{k_0}(\Theta), \mathcal{B}(\mathcal{E}_{k_0}(\Theta)))$ a prior distribution Π and obtains the posterior distribution $\Pi(dG|X^1_{[N]}, \ldots, X^m_{[N]})$ by Bayes' rule, where $\mathcal{B}(\mathcal{E}_{k_0}(\Theta))$ is the Borel sigma algebra w.r.t. D_1 distance. In this section we study the asymptotic behavior of this posterior distribution as the amount of data $m \times N$ tend to infinity. Later in Section 2.6.2 we extend the posterior contraction theory to the more realistic setting where the m sequences are of variable lengths.

Suppose throughout this section, the probability family $\{P_\theta\}_{\theta\in\Theta}$ has density $\{f(x|\theta)\}_{\theta\in\Theta}$ w.r.t. a σ-finite dominating measure μ on $\mathfrak{X}$; then $P_{G,N}$ for $G = \sum_{i=1}^{k_0} p_i\delta_{\theta_i}$ has density w.r.t. to μ:

$$p_{G,N}(\bar{x}) = \sum_{i=1}^{k_0} p_i \prod_{j=1}^{N} f(x_j|\theta_i), \quad \text{for } \bar{x} = (x_1, x_2, \cdots, x_N) \in \mathfrak{X}^N.$$

Then the density of $X^i_{[N]}$ conditioned on G is $p_{G,N}(\cdot)$. Since Θ as a subset of $\mathbb{R}^q$ is separable, $\mathcal{E}_{k_0}(\Theta)$ is separable. Moreover, suppose the map $\theta \mapsto P_\theta$ from $(\Theta, \|\cdot\|_2)$ to $(\{P_\theta\}_{\theta\in\Theta}, h)$ is continuous. Then the map from $(\mathcal{E}_{k_0}(\Theta), D_1) \to (p_{G,N}, h)$ is also continuous by Lemma 2.7.2. Then by [AK06] Lemma 4.51, $(x, G) \mapsto p_{G,N}(x)$ is measurable for each N. Thus the posterior distribution (a version of regular conditional distribution) is the random measure given by

$$\Pi(B|X^1_{[N]}, X^2_{[N]}, \ldots, X^m_{[N]}) = \frac{\int_B \prod_{i=1}^m p_{G,N}(X^i_{[N]}) d\Pi(G)}{\int_{\mathcal{E}_{k_0}(\Theta)} \prod_{i=1}^m p_{G,N}(X^i_{[N]}) d\Pi(G)}$$

for any Borel measurable subset of $B \subset \mathcal{E}_{k_0}(\Theta)$. For further details of why the last quantity is a valid posterior distribution, we refer to Section 1.3 in [GvdV17]. It is also customary to express the above Bayesian model in the following hierarchical fashion:

$$\begin{aligned} &G \sim \Pi, \quad \theta_1, \theta_2, \cdots, \theta_m | G \overset{i.i.d.}{\sim} G \\ &X_{i1}, X_{i2}, \cdots, X_{iN} | \theta_i \overset{i.i.d.}{\sim} f(x|\theta_i) \quad \text{for } i = 1, \cdots, m. \end{aligned}$$

As above, the m data sequences are denoted by $X^i_{[N]} = (X_{i1}, X_{i2}, \cdots, X_{iN}) \in \mathfrak{X}^N$ for $i = 1, 2, \cdots, m$.

The following assumptions are required for the main theorem of this section.

(B1) (Prior assumption) Suppose there is prior measure Π_θ on $\Theta_1 \subset \Theta$ with its Borel sigma algebra possessing a density w.r.t. Lebesgue measure that is bounded away from zero and infinity, where Θ_1 is a compact subset of Θ. Suppose there is a prior measure Π_p on k_0-probability simplex possessing a density w.r.t. Lebesgue measure on $\mathbb{R}^{k_0-1}$ that is bounded away from zero and infinity. Then $\Pi_p \times \Pi_\theta^{k_0}$ is a measure on $\{((p_1, \theta_1), \ldots, (p_{k_0}, \theta_{k_0}))|p_i \geq 0, \theta_i \in \Theta_1, \sum_{i=1}^{k_0} p_i = 1\}$, which induce a measure on $\mathcal{E}_{k_0}(\Theta_1)$, identified as the quotient space of preceding space by the equivalence relationship of permutation invariance.[3] Here the prior Π is generated by such mechanism with independent Π_p and Π_θ and the support Θ_1 of Π_θ is such that $G_0 \in \mathcal{E}_{k_0}(\Theta_1)$.

(B2) (Identifiability at truth) There exists $n_0 \geq 1$ such that map $G \mapsto P_{G,n_0}$ is identifiable at G_0 on $\bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1))$, i.e. for any $G \in \bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1))\backslash\{G_0\}$, $P_{G,n_0} \neq P_{G_0,n_0}$.

(B3) (Kernel assumption) Suppose $K(f(x|\theta_1), f(x|\theta_2)) \leq L_1\|\theta_1 - \theta_2\|_2^{\alpha_0}$ for some $\alpha_0 > 0$ and some $L_1 > 0$. Suppose $h(f(x|\theta_1), f(x|\theta_2)) \leq L_2\|\theta_1 - \theta_2\|_2^{\beta_0}$ for some $\beta_0 > 0$ and some $L_2 > 0$. Here θ_1, θ_2 are any distinct elements in Θ_1.

[3] Rigorously speaking, $\mathcal{E}_{k_0}(\Theta_1)$ is only a proper subspace of the quotient space since a point in the quotient space might have $\theta_i = \theta_j$. But the set of such points have probability zero under the induced measure.

Remark 2.6.1. (B1) on the boundedness of the support Θ_1 of prior distribution Π_θ is a standard assumption so as to obtain concrete rates of posterior contraction for the data population's distribution P_G (cf. [GvdV17]). Moreover, the compactness of Θ_1 is required in order to transfer the convergence rates of P_G into that of G (cf. Lemma 2.5.5).

(B2) is a necessary condition since otherwise there is no way to recover G_0. The n_0 can be taken to be $n_0(G_0)$ defined in (2.25) when $\Theta_1 \subset \Theta^\circ$. In this case, due to Proposition 2.5.1 f), it can be implied provided (2.24) is satisfied for every $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$, which holds for all full rank exponential family specified in Corollary 2.5.8, or for more general family by checking Theorem 2.5.14 for every $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$.

Note that (B3) do imply some implicit constraints on α_0 and β_0. Specifically, if (2.11) holds for some $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$ and (B3) holds, then $\beta_0 \leq 1$ and $\alpha_0 \leq 2$. Indeed, for any sequence $G_\ell = \sum_{i=2}^{k_0} p_i^0 \delta_{\theta_i^0} + p_1^0 \delta_{\theta_1^\ell} \in \mathcal{E}_{k_0}(\Theta) \backslash \{G_0\}$ converges to $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$, by (2.11), Lemma 2.11.2 with $N = 1$ and (B3), for large ℓ

$$\begin{aligned} C(G_0)\|\theta_1^\ell - \theta_1^0\|_2 = C(G_0) D_1(G_\ell, G_0) &\leq V(P_{G_\ell}, P_{G_0}) \\ &\leq V(f(x|\theta_1^\ell), f(x|\theta_1^0)) \leq h(f(x|\theta_1^\ell), f(x|\theta_1^0)) \leq L_2 \|\theta_1^\ell - \theta_1^0\|_2^{\beta_0}, \quad (2.34) \end{aligned}$$

which implies $\beta_0 \leq 1$ if divide both sides by $\|\theta_1^\ell - \theta_1^0\|_2$ and let $\ell \to \infty$. In the preceding display

$$C(G_0) = \frac{1}{2} \liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(p_G, p_{G_0})}{D_1(G, G_0)} > 0.$$

By (2.34) and Pinsker's inequality, for large ℓ

$$C(G_0)\|\theta_1^\ell - \theta_1^0\|_2 \leq V(f(x|\theta_1^\ell), f(x|\theta_1^0)) \leq \sqrt{\frac{1}{2} K(f(x|\theta_1^0), f(x|\theta_1^\ell))} \leq \sqrt{\frac{1}{2} L_1 \|\theta_1^\ell - \theta_1^0\|_2^{\alpha_0}},$$

which implies $\alpha_0 \leq 2$ if divide both sides by $\|\theta_1^\ell - \theta_1^0\|_2$ and let $\ell \to \infty$. The same conclusion holds if one replace (2.11) with (2.22) by an analogous argument. □

Theorem 2.6.2. *Fix $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$. Suppose (B1), (B2) and (B3) hold. Suppose additionally* (2.22) *holds. Let $n_1(G_0)$ be given by* (2.25)*.*

a) There exist a constant $C(G_0) > 0$ such that as $m \to \infty$ while fixing $N \geq n_1(G_0) \vee n_0$, for every $\bar{M}_m \to \infty$, there holds

$$\Pi\bigg(G \in \mathcal{E}_{k_0}(\Theta_1) : D_N(G, G_0) \geq C(G_0) \bar{M}_m \sqrt{\frac{\ln(mN)}{m}} \bigg| X_{[N]}^1, X_{[N]}^2, \ldots, X_{[N]}^m \bigg) \to 0$$

in $\bigotimes^m P_{G_0,N}$-probability as $m \to \infty$.

b) If in addition, (2.11) *is satisfied. Then the claim in part a) holds with* $n_1(G_0) = 1$.

In the above statement, note that the constant $C(G_0)$ also depends on Θ_1, k_0, q, upper and lower bounds of the densities of Π_θ, Π_p and the density family $f(x|\theta)$ (including α_0, β_0, L_1, L_2 etc). All such dependence are suppressed for the sake of clean presentation; it is the dependence on G_0 and the *independence* of m, N that we want to emphasize. Further remarks of Theorem 2.6.2 is available as in Remark 2.6.6 with $\bar{N}_m$ replaced by N. These remarks are deferred to the next section, where we establish posterior contraction in a more general setting.

As already discussed in the Remark 2.6.1, the condition (B2) is satisfied for full rank exponential families of kernels. Indeed (B3) can also be verified for for full rank exponential families and hence we have the following corollary from Theorem 2.6.2.

Corollary 2.6.3. *Consider a full rank exponential family for kernel P_θ specified as in Corollary 2.5.8 and assume all the requirements there are met. Fix $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$. Suppose that* $(B1)$ *holds with $\Theta_1 \subset \Theta^\circ$.*

a) There exist a constant $C(G_0) > 0$ such that as $m \to \infty$ while fixing $N \geq n_1(G_0) \vee n_0(G_0)$, for every $\bar{M}_m \to \infty$, there holds

$$\Pi\left(G \in \mathcal{E}_{k_0}(\Theta_1) : D_N(G, G_0) \geq C(G_0)\bar{M}_m \sqrt{\frac{\ln(mN)}{m}} \middle| X^1_{[N]}, X^2_{[N]}, \ldots, X^m_{[N]} \right) \to 0$$

in $\bigotimes^m P_{G_0,N}$-probability as $m \to \infty$.

b) If in addition, (2.11) *is satisfied. Then the claim in part a) holds with* $n_1(G_0) = 1$.

Proof: By Corollary 2.5.8 and Proposition 2.5.1 f), (2.22) holds and $n_0(G_0) < \infty$. Moreover since $\Theta_1 \subset \Theta^\circ$ (B2) holds with n_0 there be $n_0(G_0) < \infty$. Moreover, by easy calculations

$$|K(f(x|\theta_1), f(x|\theta_2))| = |\langle \theta_1 - \theta_2, \mathbb{E}_{\theta_1} Tx \rangle - (B(\theta_1) - B(\theta_2))| \leq L_1(\Theta_1)\|\theta_1 - \theta_2\|_2.$$

By changing to its canonical parametrization and appeal to Lemma 2.13.2 b),

$$|h(f(x|\theta_1) - h(f(x|\theta_2)))| \leq L_2(\Theta_1)\|\theta_1 - \theta_2\|_2.$$

Here $L_1(\Theta_1)$ and $L_2(\Theta_1)$ are constants that depend on Θ_1. In summary (B3) is satisfied. Then the conclusions are obtained by applying Theorem 2.6.2. □

We end this subsection by applying the Corollary 2.6.3 to the Bernoulli kernel and two-paramters Gamma kernel.

Example 2.6.4 (Posterior contraction for weakly identifiable kernels: Bernoulli and Gamma)**.** Fix $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)$. Suppose that $(B1)$ holds with $\Theta_1 \subset \Theta^\circ$.

Consider the Bernoulli kernel as in Example 2.5.10. As already shown in Example 2.5.10, $n_1(G_0) = n_0(G_0) = 2k_0 - 1$. Then the conclusion a) in Corollary 2.6.3 holds for any $N \geq 2k_0 - 1$.

Consider the Gamma kernel as in Example 2.4.18, 2.5.12. As already shown in Example 2.5.12, $n_1(G_0) = 2$ when $G_0 \in \mathcal{G}$ and $n_1(G_0) = 1$ when $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ)\backslash\mathcal{G}$; $n_0(G_0) = 1$. Then the conclusion a) in Corollary 2.6.3 holds for any $N \geq 2$. Moreover, the same conclusion holds with $N \geq 1$ if additional information $G_0 \notin \mathcal{G}$ is known. □

2.6.2 Data are variable-length exchangeable sequences

Now we turn to a realistic setting where the m observed exchangeable sequences are of variable lengths. For $i = 1, 2, \cdots, m$, denote $X^i_{[N_i]} = (X_{i1}, X_{i2}, \cdots, X_{iN_i}) \in \mathfrak{X}^{N_i}$, where N_i is the length of $X^i_{[N_i]}$ sequence. Consider the hierarchical model:

$$\begin{aligned} &G \sim \Pi, \quad \theta_1, \theta_2, \cdots, \theta_m | G \overset{i.i.d.}{\sim} G \\ &X_{i1}, X_{i2}, \cdots, X_{iN_i} | \theta_i \overset{i.i.d.}{\sim} f(x|\theta_i) \quad \text{for } i = 1, \cdots, m. \end{aligned}$$

As in Section 2.6.1, the probability family $\{P_\theta\}_{\theta\in\Theta}$ has density $\{f(x|\theta)\}_{\theta\in\Theta}$ w.r.t. a σ-finite measure μ on $\mathfrak{X}$; then P_{G,N_i} for $G = \sum_{i=1}^{k_0} p_i\delta_{\theta_i}$ has density w.r.t. to μ:

$$p_{G,N_i}(\bar{x}) = \sum_{i=1}^{k_0} p_i \prod_{j=1}^{N_i} f(x_j|\theta_i), \quad \text{for } \bar{x} = (x_1, x_2, \cdots, x_{N_i}) \in \mathfrak{X}^{N_i}.$$

Then the density of $X^i_{[N_i]}$ conditioned on G is $p_{G,N_i}(\cdot)$. Thus, given the m data sequences $X^1_{[N_1]}, \ldots, X^m_{[N_m]}$, the posterior distribution of mixing measure G is given by

$$\Pi(B|X^1_{[N_1]}, \ldots, X^m_{[N_m]}) = \frac{\int_B \prod_{i=1}^m p_{G,N_i}(X^i_{[N_i]}) d\Pi(G)}{\int_{\mathcal{E}_{k_0}(\Theta)} \prod_{i=1}^m p_{G,N_i}(X^i_{[N_i]}) d\Pi(G)}, \text{ for } B \text{ measurable subset of } \mathcal{E}_{k_0}(\Theta).$$

An useful quantity is the average sequence length

$$\bar{N}_m = \frac{1}{m}\sum_{i=1}^m N_i.$$

In fact, the posterior contraction theorem will be characterized in terms of distance $D_{\bar{N}_m}(\cdot,\cdot)$, which extends the original notion of distance $D_N(\cdot,\cdot)$ by allowing real-valued weight $\bar{N}_m$. Specifically,

for any real $r > 0$, for $G = \sum_{i=1}^{k_0} p_i \delta_{\theta_i} \in \mathcal{E}_{k_0}(\Theta)$ and $G' = \sum_{i=1}^{k_0} p'_i \delta_{\theta'_i} \in \mathcal{E}_{k_0}(\Theta)$ define

$$D_r(G, G') = \min_{\tau \in S_{k_0}} \sum_{i=1}^{k_0} (\sqrt{r}\|\theta_{\tau(i)} - \theta'_i\|_2 + |p_{\tau(i)} - p'_i|),$$

where S_{k_0} denotes the set of all permutations on the set $[k_0]$.

Theorem 2.6.5. *Fix $G_0 \in \mathcal{E}_{k_0}(\Theta)$. Suppose (B1), (B2), and (B3) hold. Suppose additionally* (2.22) *holds. Let $n_1(G_0)$ be given by* (2.25)*.*

a) There exists some constant $C(G_0) > 0$ such that as long as $n_0 \vee n_1(G_0) \leq \min_i N_i \leq \max_i N_i < \infty$, for every $\bar{M}_m \to \infty$ there holds

$$\Pi\left(G \in \mathcal{E}_{k_0}(\Theta_1) : D_{\bar{N}_m}(G, G_0) \geq C(G_0)\bar{M}_m \sqrt{\frac{\ln(m\bar{N}_m)}{m}} \middle| X^1_{[N_1]}, \ldots, X^m_{[N_m]} \right) \to 0$$

in $\bigotimes_{i=1}^m P_{G_0, N_i}$-probability as $m \to \infty$.

b) If in addition, (2.11) *is satisfied. Then the claim in part a) holds with $n_1(G_0) = 1$.*

We make the following remarks.

Remark 2.6.6. a) In the above statement, note that the constant $C(G_0)$ also depends on Θ_1, k_0, q, upper and lower bounds of the densities of Π_θ, Π_p and the density family $f(x|\theta)$ (including α_0, β_0, L_1, L_2 etc). All such dependence are suppressed for the sake of clean presentation; it is the dependence on G_0 and the *independence* of $m, \{N_i\}_{i \geq 1}$ and $N_0 := \sup_i N_i < \infty$, that we want to emphasize. In addition, although $C(G_0)$ and hence the vanishing radius of the ball characterized by $D_{\bar{N}_m}$ does *not* depend on N_0, the rate at which the posterior probability statement concerning this ball tending to zero may depend on it.

b) Roughly speaking, the theorem produces the following posterior contraction rates. The posterior contraction toward mixing probabilities p_i^0 is of the rate $O_P((\ln(m\bar{N}_m)/m)^{1/2})$. Individual atoms θ_i^0 receive much faster posterior contract rate, which utilizes the full volume of the data set:

$$O_P(\ln(m\bar{N}_m)/m\bar{N}_m)^{1/2}) = O_P\left(\sqrt{\frac{\ln(\sum_{i=1}^m N_i)}{\sum_{i=1}^m N_i}} \right). \tag{2.35}$$

c) The distinction between the two parts of the theorem highlights the role of first-order identifiability in mixtures of N-product distributions. Under first-order identifiability, (2.11) is satisfied, so we can establish the aforementioned posterior contraction behavior for a full range of sequence length N_i's, as long as they are uniformly bounded by an arbitrary unknown

constant. When first-order identifiability is not satisfied, so (2.11) may fail to hold, the same posterior behavior can be attained only when the sequence lengths exceed certain constant depending on the true G_0.

□

2.7 Sharpness of bounds and minimax theorem

Much efforts in the previous sections, particularly Section 2.4 and Section 2.5, were devoted to establishing so-called inverse bounds for mixtures of product distributions of exchangeable sequences. These are lower bounds of distances between a pair of distributions $(P_{G_0,N}, P_{G,N})$ in terms of distance $D_N(G_0, G)$ between corresponding de Finetti's mixing measures (G_0, G). The distance $D_N(G_0, G)$ brings out the role of the sample size N of exchangeable sequences. Under suitable identifiability and regularity conditions we established the convergence of certain mixing parameters with a rate proportional to $N^{-1/2}$. In this section, we shall examine the sharpness of the inverse bounds obtained, by presenting suitable opposing upper bounds. These relatively easy upper bounds are also useful in establishing minimax theorems for the estimation of mixing measures.

2.7.1 Sharpness of inverse bounds

Inverse bounds of the form (2.22) hold only under some identifiability conditions, while the following upper bound holds generally and is much easier to show.

Lemma 2.7.1. *Let $k_0 \geq 2$ and fix $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta)$. Then for any $N \geq 1$*

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)} \leq \liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_1(G, G_0)} \leq \min_{1 \leq i < j \leq k_0} \frac{1}{2} V(\bigotimes^N P_{\theta_i^0}, \bigotimes^N P_{\theta_j^0}) \leq \frac{1}{2}.$$

Proof: Consider $G_\ell = \sum_{i=1}^{k_0} p_i^\ell \delta_{\theta_i^0}$ with $p_i^\ell = p_i^0$ for $3 \leq \ell \leq k_0$ and $p_1^\ell = p_1^0 + \frac{1}{\ell}$, $p_2^\ell = p_2^0 - \frac{1}{\ell}$. Then for sufficiently large ℓ, $p_1^\ell, p_2^\ell \in (0, 1)$ and hence $G_\ell \in \mathcal{E}_{k_0}(\Theta) \backslash \{G_0\}$ and satisfies $D_N(G_\ell, G_0) = D_1(G_\ell, G_0) = 2/\ell$. Thus for sufficiently large ℓ,

$$\frac{V(P_{G,N}, P_{G_0,N})}{D_1(G, G_0)} = \frac{\ell}{2} \sup_{A \in \mathcal{A}} \left| \frac{1}{\ell} \bigotimes^N P_{\theta_1^0}(A) - \frac{1}{\ell} \bigotimes^N P_{\theta_2^0}(A) \right| = \frac{1}{2} V(\bigotimes^N P_{\theta_1^0}, \bigotimes^N P_{\theta_2^0}) \leq \frac{1}{2}.$$

The proof is then complete by observing the above analysis indeed holds for any pair θ_i^0, θ_j^0 instead of θ_1^0, θ_2^0. □

The next lemma establishes a upper bound for Hellinger distance of two mixture of product

measures by Hellinger distance of individual components. Such result is useful later in Lemma 2.7.3 and Theorem 2.7.5. A similar result on variation distance is Lemma 2.11.2.

Lemma 2.7.2. *For any* $G = \sum_{i=1}^{k_0} p_i \delta_{\theta_i}$ *and* $G' = \sum_{i=1}^{k_0} p'_i \delta_{\theta'_i}$,

$$h(P_{G,N}, P_{G',N}) \leq \min_{\tau} \left(\sqrt{N} \max_{1 \leq i \leq k_0} h\left(P_{\theta_i}, P_{\theta'_{\tau(i)}}\right) + \sqrt{\frac{1}{2} \sum_{i=1}^{k_0} \left| p_i - p'_{\tau(i)} \right|} \right),$$

where the minimum is taken over all τ *in the permutation group* S_{k_0}.

The inverse bounds expressed by Eq. (2.22) are optimal as far as the role of N in D_N is concerned. This is made precise by the following result.

Lemma 2.7.3 (Optimality of $\sqrt{N}$)**.** *Fix* $G_0 = \sum_{i=1}^{k_0} p_i \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta^\circ)$. *Suppose there exists* $j \in [k_0]$ *such that* $\liminf_{\theta \to \theta_j^0} \frac{h(P_\theta, P_{\theta_j^0})}{\|\theta - \theta_j^0\|_2} < \infty$. *Then for* $\psi(N)$ *such that* $\frac{\psi(N)}{N} \to \infty$,

$$\limsup_{N \to \infty} \liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{h(P_{G,N}, P_{G_0,N})}{D_{\psi(N)}(G, G_0)} = 0.$$

A slightly curious and pedantic way to gauge the meaning of the double infimum limiting arguments in the inverse bound (2.22), is to express its claim as follows:

$$0 < \liminf_{N \to \infty} \liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G \in \mathcal{E}_{k_0}(\Xi)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)} = \lim_{k \to \infty} \inf_{N \geq k} \lim_{\epsilon \to 0} \inf_{G \in B_{W_1}(G_0, \epsilon) \setminus \{G_0\}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)},$$

where $B_{W_1}(G_0, R) \subset \mathcal{E}_{k_0}(\Theta)$ is defined in (2.10). It is possible to alter the order of the four operations and consider the resulting outcome. The following lemma shows the last display is the only order to possibly obtain a positive outcome.

Lemma 2.7.4. *a)*

$$\lim_{k \to \infty} \lim_{\epsilon \to 0} \inf_{N \geq k} \inf_{G \in B_{W_1}(G_0, \epsilon) \setminus \{G_0\}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)}$$
$$= \lim_{k \to \infty} \lim_{\epsilon \to 0} \inf_{G \in B_{W_1}(G_0, \epsilon) \setminus \{G_0\}} \inf_{N \geq k} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)} = 0$$

b)

$$\lim_{\epsilon\to 0}\lim_{k\to\infty}\inf_{N\geq k}\inf_{G\in B_{W_1}(G_0,\epsilon)\setminus\{G_0\}}\frac{V(P_{G,N},P_{G_0,N})}{D_N(G,G_0)}$$
$$=\lim_{\epsilon\to 0}\lim_{k\to\infty}\inf_{G\in B_{W_1}(G_0,\epsilon)\setminus\{G_0\}}\inf_{N\geq k}\frac{V(P_{G,N},P_{G_0,N})}{D_N(G,G_0)}=0$$

c)

$$\lim_{\epsilon\to 0}\inf_{G\in B_{W_1}(G_0,\epsilon)\setminus\{G_0\}}\lim_{k\to\infty}\inf_{N\geq k}\frac{V(P_{G,N},P_{G_0,N})}{D_N(G,G_0)}=0.$$

Proof: The claims follow from

$$\inf_{N\geq k}\inf_{G\in B_{W_1}(G_0,\epsilon)\setminus\{G_0\}}\frac{V(P_{G,N},P_{G_0,N})}{D_N(G,G_0)}=\inf_{G\in B_{W_1}(G_0,\epsilon)\setminus\{G_0\}}\inf_{N\geq k}\frac{V(P_{G,N},P_{G_0,N})}{D_N(G,G_0)}$$

and

$$\inf_{N\geq k}\frac{V(P_{G,N},P_{G_0,N})}{D_N(G,G_0)}\leq\inf_{N\geq k}\frac{1}{D_N(G,G_0)}=0.$$

□

2.7.2 Minimax lower bounds

Given $G=\sum_{i=1}^{k_0}p_i\delta_{\theta_i}\in\mathcal{E}_{k_0}(\Theta)$ and $G'=\sum_{i=1}^{k_0}p'_i\delta_{\theta'_i}\in\mathcal{E}_{k_0}(\Theta)$, define additional notions of distances[4]

$$d_{\boldsymbol{\theta}}(G',G):=\min_{\tau\in S_{k_0}}\sum_{i=1}^{k_0}\|\theta'_{\tau(i)}-\theta_i\|_2 \tag{2.36}$$

$$d_{\boldsymbol{p}}(G',G):=\min_{\tau\in S_{k_0}}\sum_{i=1}^{k_0}|p'_{\tau(i)}-p_i|. \tag{2.37}$$

These two notions of distance are pseudometrics on the space of measures $\mathcal{E}_{k_0}(\Theta)$, i.e., they share the same properties as a metric except that allow the distance between two different points may be zero. $d_{\boldsymbol{\theta}}(G',G)$ focus on the distance between atoms of two mixing measure; while $d_{\boldsymbol{p}}(G',G)$ focus on the mixing probabilities of the two mixing measures. It is clear that

$$D_1(G,G')\geq d_{\boldsymbol{\theta}}(G,G')+d_{\boldsymbol{p}}(G,G'). \tag{2.38}$$

We proceed to present minimax lower bounds for any sequence of estimates $\hat{G}$, which are measurable functions of $X^1_{[N]},\ldots,X^m_{[N]}$. The minimax bounds are stated in terms of the aforementioned (pseudo-)metrics $d_{\boldsymbol{p}}$ and $d_{\boldsymbol{\theta}}$, as well as the usual metric D_1 studied.

[4]Notice that we denote d_θ to be a distance on Θ in Section 2.3. Here the $d_{\boldsymbol{\theta}}$ with bold subscript is on $\mathcal{E}_{k_0}(\Theta)$.

Theorem 2.7.5 (Minimax Lower Bound). *a) Suppose there exists $\theta_0 \in \Theta^\circ$ and $\beta_0 > 0$ such that $\limsup_{\theta\to\theta_0} \frac{h\left(P_\theta, P_{\theta_0}\right)}{\|\theta-\theta_0\|_2^{\beta_0}} < \infty$. Moreover, suppose there exists a set of distinct $k_0 - 1$ points $\{\theta_i\}_{i=1}^{k_0-1} \subset \Theta\backslash\{\theta_0\}$ satisfying $\rho_1 = \min_{0\le i<j\le k_0-1} h(P_{\theta_i}, P_{\theta_j}) > 0$. Then*

$$\inf_{\hat{G}\in\mathcal{E}_{k_0}(\Theta)} \sup_{G\in\mathcal{E}_{k_0}(\Theta)} \mathbb{E}_{\bigotimes^m P_{G,N}} d_{\boldsymbol{\theta}}(\hat{G}, G) \ge C(\beta_0) \left(\frac{1}{\sqrt{m}\sqrt{N}}\right)^{\frac{1}{\beta_0}}.$$

b) Let $k_0 \ge 2$.

$$\inf_{\hat{G}\in\mathcal{E}_{k_0}(\Theta)} \sup_{G\in\mathcal{E}_{k_0}(\Theta)} \mathbb{E}_{\bigotimes^m P_{G,N}} d_{\boldsymbol{p}}(\hat{G}, G) \ge C(k_0)\frac{1}{m},$$

c) Let $k_0 \ge 2$. Provided that the assumptions of part (a) holds. Then,

$$\inf_{\hat{G}\in\mathcal{E}_{k_0}(\Theta)} \sup_{G\in\mathcal{E}_{k_0}(\Theta)} \mathbb{E}_{\bigotimes^m P_{G,N}} D_1(\hat{G}, G) \ge C(\beta_0) \left(\frac{1}{\sqrt{m}\sqrt{N}}\right)^{\frac{1}{\beta_0}} + C(k_0)\frac{1}{m}.$$

In all three bounds the infimum is taken for all $\hat{G}$ measurable functions of $X^1_{[N]}, \ldots, X^m_{[N]}$.

Remark 2.7.6. a) The condition that there exists a set of distinct $k_0 - 1$ points $\{\theta_i\}_{i=1}^{k_0-1} \subset \Theta\backslash\{\theta_0\}$ satisfying $\rho_1 = \min_{0\le i<j\le k_0-1} h(P_{\theta_i}, P_{\theta_j}) > 0$ immediately follows from the injectivity of the map $\theta \mapsto P_\theta$ (recall that this condition is assumed throughout the chapter).

b) The condition that there exists $\theta_0 \in \Theta^\circ$ and $\beta_0 > 0$ such that $\limsup_{\theta\to\theta_0} \frac{h\left(P_\theta, P_{\theta_0}\right)}{\|\theta-\theta_0\|_2^{\beta_0}} < \infty$ holds for most probability kernels considered in practice. For example, it is satisfied with $\beta_0 = 1$ for all full rank exponential families of distribution in their canonical form as shown by Lemma 2.13.2. It can then be shown that this condition with $\beta_0 = 1$ is also satisfied by full rank exponential families in general form specified in Corollary 2.5.8. Notice that the same remark applies to the condition in Lemma 2.7.3.

c) If conditions of Theorem 2.7.5 a) hold with $\beta_0 = 1$, then

$$\inf_{\hat{G}\in\mathcal{E}_{k_0}(\Theta)} \sup_{G\in\mathcal{E}_{k_0}(\Theta)} \mathbb{E}_{\bigotimes^m P_{G,N}} d_{\boldsymbol{\theta}}(\hat{G}, G) \ge \frac{C}{\sqrt{m}\sqrt{N}}.$$

That is, the convergence rate of the best possible estimator for the worst scenario is at least $\frac{1}{\sqrt{m}\sqrt{N}}$. Recall that Theorem 2.6.2 implied that the convergence rate of the atoms is $O_P(\sqrt{\frac{\ln(mN)}{mN}})$, which is obtained by replacing $\bar{N}_m$ with N in (2.35). It is worth noting that while the minimax rate seems to match the posterior contraction rate of the atoms except for a logarithmic factor, such comparison is not very meaningful as pointwise posterior contraction

bounds and (local) minimax lower bounds are generally not considered to be compatible. In addition, in the posterior contraction theorems presented in the previous section, the truth G_0 is fixed and the hidden constant $O_P(\sqrt{\frac{\ln(mN)}{mN}})$ depends on G_0, which is clearly not the case in the above results obtained under the minimax framework. In short, we do not claim that the Bayesian procedure described in the previous section is optimal in the minimax sense; nor do we claim that the bounds given in Theorem 2.7.5 are sharp (i.e., achievable by some statistical procedure). □

2.8 Hierarchical model: kernel P_θ is itself a mixture distribution

In this section we apply Theorem 2.5.14 to the cases where P_θ itself is a rather complex object: a finite mixture of distributions. Combining this kernel with a discrete mixing measure $G \in \mathcal{E}_{k_0}(\Theta)$, the resulting P_G represents a mixture of finite mixtures of distributions, while $P_{G,N}$ becomes a k_0-mixture of N-products of finite mixtures of distributions. These recursively defined objects represent a formidable and popular device in the statistical modeling world: the world of hierarchical models. In this section we shall illustrate Theorem 2.5.14 on only two examples of such models. However, the tools required for these applications are quite general, chief among them are bounds on relevant oscillatory integrals for suitable statistical maps T. We shall first describe such tools in subsection 2.8.1 and then proceed to applying Theorem 2.5.14 to the case P_θ is a k-component Gaussian location mixture as introduced in Example 2.5.20. Finally, subsection 2.8.4 studies the case P_θ is a mixture of Dirichlet processes as introduced in Example 2.5.21.

2.8.1 Bounds on oscillatory integrals

A key condition in Theorem 2.5.14, namely condition (A3), is reduced to the L^r integrability of certain oscillatory integrals:

$$\left\| \int_{\mathfrak{X}} e^{\boldsymbol{i}\zeta^T Tx} f(x) dx \right\|_{L^r(\mathbb{R}^s)} \tag{2.39}$$

for a broad class of functions $f : \mathfrak{X} \to \mathbb{R}$ and multi-dimensional maps $T : \mathfrak{X} \to \mathbb{R}^s$. When $\mathfrak{X} = \mathbb{R}^d$, the oscillatory integral $\int_{\mathfrak{X}} e^{\boldsymbol{i}\zeta^T Tx} f(x) dx$ is also known as the Fourier transform of measures supported on curves or surfaces; bounds for such quantities are important topics in harmonic analysis and geometric analysis. We refer to [BGG$^+$07] and the textbook [SM93] (Chapter 8) for further details and broader contexts. Despite there are many existing results, such results are typically established when $f(x)$ is supported on a compact interval or is smooth, i.e. f has derivative of arbitrary orders. We shall develop an upper bound on (2.39) for our purposes to verify the integrability condition in (A3) for a broad class of f, which is usually satisfied for probability density functions.

We will start by stating the following fact, which provides an useful estimate for oscillatory integrals of the form $\int e^{i\lambda\phi(x)}\psi(x)dx$, where function ϕ is called the phase, and function ψ the amplitude.

Lemma 2.8.1 (van der Corput's Lemma)**.** *Suppose $\phi(x) \in C^\infty(a,b)$, and that $|\phi^{(k)}(x)| \geq 1$ for all $x \in (a,b)$. Let $\psi(x)$ be absolute continuous on $[a,b]$. Then*

$$\left|\int_{[a,b]} e^{i\lambda\phi(x)}\psi(x)dx\right| \leq c_k\lambda^{-\frac{1}{k}}\left[|\psi(b)| + \int_{[a,b]}|\psi'(x)|dx\right]$$

and

$$\left|\int_{[a,b]} e^{i\lambda\phi(x)}\psi(x)dx\right| \leq c_k\lambda^{-\frac{1}{k}}\left[|\psi(a)| + \int_{[a,b]}|\psi'(x)|dx\right]$$

hold when:

1. *$k \geq 2$, or*
2. *$k = 1$ and $\phi'(x)$ is monotonic.*

The constant c_k is independent of ϕ, ψ, λ and the interval $[a,b]$.

Proof: See the Corollary on Page 334 of [SM93] for the proof of the first display; even in its original version in this reference, ψ is assumed to be C^∞ but its proof only needs ψ to be absolute continuous on $[a,b]$. The second display follows by applying the first display to $\tilde{\psi}(x) = \psi(a+b-x)$. □

It can be observed from Lemma 2.8.1 the condition on derivatives of the phase function plays a crucial role. For our purpose the phase function will be supplied by use of monomial map T. Hence, the following technical lemma will be needed.

Lemma 2.8.2. *Let $A(x) \in \mathbb{R}^{d\times d}$ with entries $A_{\alpha\beta}(x) = 0$ for $\alpha > \beta$ and $A_{\alpha\beta}(x) = \frac{j_\beta!}{(j_\beta - j_\alpha)!}x^{j_\beta - j_\alpha}$ for $1 \leq \alpha \leq \beta \leq d$, where $1 \leq j_1 < \ldots < j_d$ are given. Let $S_{min}(A(x))$ be the smallest singular value of $A(x)$. Then $S_{min}(A(x)) \geq c_3 \max\{1,|x|\}^{-(j_d - j_1)(d-1)}$, where c_3 is a constant that depends only on $d, j_1, \ldots, j_d$.*

The following lemma provides a crucial uniform bound on oscillatory integrals given by a phase given by monomial map T.

Lemma 2.8.3. *Let $T : \mathbb{R} \to \mathbb{R}^d$ defined by $Tx = (x^{j_1}, x^{j_2}, \ldots, x^{j_d})^T$ with $1 \leq j_1 < j_2 < \ldots j_d$. Consider a bounded non-negative function $f(x)$ that is differentiable on $\mathbb{R}\backslash\{b_i\}_{i=1}^{\ell}$, where $b_1 < b_2 < \ldots < b_\ell$ with ℓ a finite number. The derivative $f'(x) \in L^1(\mathbb{R})$ and is continuous when exists.*

Moreover, $f(x)$ and $|x|^{\alpha_1} f(x)$ are both increasing when $x < -c_1$ and decreasing when $x > c_1$ for some $c_1 \geq \max\{|b_1|, |b_\ell|\}$, where $\alpha_1 = (j_d - j_1)(d-1)/j_1$. Then for $\lambda > 1$,

$$\begin{aligned}
&\sup_{w \in S^{d-1}} \left| \int_{\mathbb{R}} \exp(\boldsymbol{i}\lambda w^T T x) f(x) dx \right| \\
\leq & C(d, j_1, \ldots, j_d) \lambda^{-\frac{1}{j_d}} (c_1 + 2)^{\alpha_1} \Big(\| |x|^{\alpha_1} f(x) \|_{L^1(\mathbb{R})} + (\ell + 1) \|f\|_{L^\infty(\mathbb{R})} + \\
& \| (|x|^{\alpha_1} + 1) f'(x) \|_{L^1(\mathbb{R})} \Big),
\end{aligned}$$

where $C(d, j_1, j_2, \ldots, j_d)$ is a constant that only depends on its parameters.

Applying Lemma 2.8.3 we can obtain the following bound for the oscillatory integral in question.

Lemma 2.8.4. *Let T and f satisfy the same conditions as in Lemma 2.8.3. Define $g(\zeta) = \int_{\mathbb{R}} e^{\boldsymbol{i}\zeta^T T x} f(x) dx$ for $\zeta \in \mathbb{R}^d$. Then for $r > d j_d$,*

$$\begin{aligned}
&\|g(\zeta)\|_{L^r(\mathbb{R}^d)} \\
\leq & C(r, d, j_1, \ldots, j_d)(c_1 + 2)^{\alpha_1} \left(\| |x|^{\alpha_1} f(x) \|_{L^1(\mathbb{R})} + (\ell + 1) \|f\|_{L^\infty(\mathbb{R})} + \right. \\
& \left. \| (|x|^{\alpha_1} + 1) f'(x) \|_{L^1(\mathbb{R})} + \|f\|_{L^1(\mathbb{R})} \right)
\end{aligned}$$

where $C(r, d, j_1, \ldots, j_d)$ is a constant that depends on its parameters.

2.8.2 Kernel P_θ is a mixture of Gaussian distributions

This and the next subsection are devoted to the application of Theorem 2.5.14 to case kernel P_θ is a mixture of k Gaussian distributions. Let

$$\Theta = \{\theta = (\pi_1, \pi_2, \ldots, \pi_{k-1}, \mu_1, \mu_2, \ldots, \mu_k) \in \mathbb{R}^{2k-1} | 0 < \pi_i < 1,\ \forall i;\ \mu_i < \mu_j,\ \forall 1 \leq i < j \leq k\}$$

and P_θ w.r.t. Lebesgue measure on $\mathbb{R}$ has probability density

$$f(x|\theta) = \sum_{i=1}^{k} \pi_i f_{\mathcal{N}}(x|\mu_i, \sigma^2) \tag{2.40}$$

where $\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$ and $f_{\mathcal{N}}(x|\mu, \sigma^2)$ is the density of $\mathcal{N}(\mu, \sigma^2)$ with σ a known constant. The goal is to recover $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ with $\theta_i^0 = (\pi_{1i}^0, \ldots, \pi_{(k-1)i}^0, \mu_{1i}^0, \ldots, \mu_{ki}^0)$ from sequences $\bar{x}$ distributed according to the mixture of product distributions, which admits the following density

w.r.t. Lebesgue measure on $\mathbb{R}^N$:

$$p_{G_0,N}(\bar{x}) = \sum_{i=1}^{k} p_i^0 \prod_{j=1}^{N} f(x_j|\theta_i^0), \text{ for } \bar{x} = (x_1^T, \ldots, x_N^T)^T \in \mathbb{R}^N,$$

where f is given by (2.40). Let us now verify that Corollary 2.5.15 with the map

$$Tx = (x, x^2, \ldots, x^{2k-1})^T$$

can be applied for this model.

The mean of TX_1 is $\lambda(\theta) \in \mathbb{R}^{2k-1}$ with its j-th entry given by

$$\lambda^{(j)}(\theta) = \mathbb{E}_\theta X_1^j = \sum_{i=1}^{k} \pi_i \mathbb{E}(\sigma Y + \mu_i)^j, \quad j = 1, \ldots, 2k-1 \tag{2.41}$$

where X_1 has density (2.40) and and Y is one dimensional standard Gaussian distribution $\mathcal{N}(0,1)$. The covariance matrix of TX_1 is $\Lambda(\theta) \in \mathbb{R}^{(2k-1)\times(2k-1)}$ with its $j\beta$ entries given by

$$\Lambda_{j\beta}(\theta) = \mathbb{E}_\theta X_1^{j+\beta} - \lambda^{(j)}(\theta)\lambda^{(\beta)}(\theta) = \sum_{i=1}^{k} \pi_i \mathbb{E}(\sigma Y + \mu_i)^{j+\beta} - \lambda^{(j)}(\theta)\lambda^{(\beta)}(\theta).$$

It follows immediately from these formulae that $\lambda(\theta)$ and $\Lambda(\theta)$ are continuous on Θ. That is, (A1) in Definition 2.5.13 is satisfied.

The characteristic function of TX_1 is

$$\phi_T(\zeta|\theta) = \mathbb{E}_\theta \exp(\boldsymbol{i}\zeta^T T X_1) = \sum_{i=1}^{k} \pi_i h(\zeta|\mu_i, \sigma) \tag{2.42}$$

where $h(\zeta|\mu,\sigma) = \mathbb{E}\exp(\boldsymbol{i}\zeta^T T(\sigma Y + \mu))$. Denote by $f_{\mathcal{N}}(x|\mu,\sigma)$ the density of $\mathcal{N}(\mu,\sigma^2)$. The verification of (A2) in Definition 2.5.13 is omitted since the essence is the same as the next four equations due to the dominated convergence theorem.

It is easy to verify by the dominated convergence theorem or Pratt's Lemma:

$$\frac{\partial h(\zeta|\mu,\sigma)}{\partial \mu} = \int_{\mathbb{R}} \exp(\boldsymbol{i}\sum_{i=1}^{k} \zeta^{(i)} x^i) \frac{\partial f_{\mathcal{N}}(x|\mu,\sigma)}{\partial \mu} dx,$$

$$\frac{\partial^2 h(\zeta|\mu,\sigma)}{\partial \mu^2} = \int_{\mathbb{R}} \exp(\boldsymbol{i}\sum_{i=1}^{k} \zeta^{(i)} x^i) \frac{\partial^2 f_{\mathcal{N}}(x|\mu,\sigma)}{\partial \mu^2} dx,$$

$$\frac{\partial h(\zeta|\mu,\sigma)}{\partial \zeta^{(j)}} = \int_{\mathbb{R}} \boldsymbol{i} x^j \exp(\boldsymbol{i} \sum_{i=1}^{k} \zeta^{(i)} x^i) f_{\mathcal{N}}(x|\mu,\sigma) dx, \quad j \in [k]$$

and

$$\frac{\partial^2 h(\zeta|\mu,\sigma)}{\partial \zeta^{(j)} \partial \mu} = \int_{\mathbb{R}} \boldsymbol{i} x^j \exp(\boldsymbol{i} \sum_{i=1}^{k} \zeta^{(i)} x^i) \frac{\partial f_{\mathcal{N}}(x|\mu,\sigma)}{\partial \mu} dx, \quad j \in [k].$$

Then

$$\left| \frac{\partial h(\zeta|\mu,\sigma)}{\partial \mu} \right| \leq \int_{\mathbb{R}} \left| \frac{\partial f_{\mathcal{N}}(x|\mu,\sigma)}{\partial \mu} \right| dx = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} \tag{2.43}$$

$$\left| \frac{\partial^2 h(\zeta|\mu,\sigma)}{\partial \mu^2} \right| \leq \int_{\mathbb{R}} \left| \frac{\partial^2 f_{\mathcal{N}}(x|\mu,\sigma)}{\partial \mu^2} \right| dx \leq \frac{2}{\sigma^2}, \tag{2.44}$$

$$\max_{j\in[k]} \left| \frac{\partial h(\zeta|\mu,\sigma)}{\partial \zeta^{(j)}} \right| \leq \max_{j\in[k]} \int_{\mathbb{R}} \left| x^j f_{\mathcal{N}}(x|\mu,\sigma) \right| dx := h_1(\mu), \tag{2.45}$$

$$\max_{j\in[k]} \left| \frac{\partial^2 h(\zeta|\mu,\sigma)}{\partial \zeta^{(j)} \partial \mu} \right| \leq \max_{j\in[k]} \int_{\mathbb{R}} \left| x^j \frac{\partial f_{\mathcal{N}}(x|\mu,\sigma)}{\partial \mu} \right| dx := h_2(\mu), \tag{2.46}$$

where $h_1(\mu)$ and $h_2(\mu)$ are continuous functions of μ, with their dependence on the constant σ suppressed. It follows that gradient w.r.t. θ is

$$\begin{aligned} &\nabla_\theta \phi_T(\zeta|\theta) \\ &= \left(h(\zeta|\mu_1,\sigma) - h(\zeta|\mu_k,\sigma), \ldots, h(\zeta|\mu_{k-1},\sigma) - h(\zeta|\mu_k,\sigma), \pi_1 \frac{\partial h(\zeta|\mu_1,\sigma)}{\partial \mu}, \ldots, \pi_k \frac{\partial h(\zeta|\mu_k,\sigma)}{\partial \mu} \right)^T \end{aligned} \tag{2.47}$$

and Hessian w.r.t. θ with ij entry for $j \geq i$ given by

$$\frac{\partial^2}{\partial \theta^{(j)} \partial \theta^{(i)}} \phi_T(\zeta|\theta) = \begin{cases} \frac{\partial h(\zeta|\mu_i,\sigma)}{\partial \mu} & i \in [k-1], j = k-1+i \\ -\frac{\partial h(\zeta|\mu_k,\sigma)}{\partial \mu} & i \in [k-1], j = 2k-1 \\ \pi_{i-(k-1)} \frac{\partial^2 h(\zeta|\mu_{i-(k-1)},\sigma)}{\partial \mu^2} & k \leq i \leq 2k-1, j = i \\ 0 & \text{otherwise} \end{cases} \tag{2.48}$$

and the lower part is symmetric to the upper part.

Then by (2.43), (2.44), (2.45), (2.46), (2.47) and (2.48), for any $i, j \in [k]$:

$$\left| \frac{\partial \phi_T(\zeta|\theta)}{\partial \theta^{(i)}} \right| \leq 2 + \sqrt{\frac{2}{\pi}} \frac{1}{\sigma},$$

$$\left| \frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \theta^{(i)} \partial \theta^{(j)}} \right| \leq \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} + \frac{2}{\sigma^2},$$

$$\left|\frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \zeta^{(j)} \partial \theta^{(i)}}\right| \leq \sum_{i=1}^{k} \left(h_1(\mu_i) + h_2(\mu_i)\right),$$

where the right hand side of the last display is a continuous function of θ since h_1 and h_2 are continuous. Hence to verify the condition (A3) of Theorem 2.5.14 it suffices to establish there exists some $r \geq 1$ such that $\int_{\mathbb{R}^{2k-1}} |\phi_T(\zeta|\theta)|^r \, d\zeta$ on Θ is upper bounded by a finite continuous function of θ.

As we have seen the verification of conditions (A1), (A2) and parts of (A3) in Definition 2.5.13 is fairly straightforward by simply checking the specific definition of the probability kernel P_θ. The remaining condition of (A3) requires results from subsection 2.8.1.

Note that $f_{\mathcal{N}}(x|\mu,\sigma)$ is differentiable everywhere and $\frac{\partial f_{\mathcal{N}}(x|\mu,\sigma)}{\partial x} \in L^1(\mathbb{R})$. Moreover α_1 in Lemma 2.8.4 for T is $4(k-1)^2$ and $f_{\mathcal{N}}(x|\mu,\sigma)$, $x^{4(k-1)^2} f_{\mathcal{N}}(x|\mu,\sigma)$ are increasing on $\left(-\infty, -\frac{|\mu|+\sqrt{\mu^2+16(k-1)^2\sigma^2}}{2}\right)$ and decreasing on $\left(\frac{|\mu|+\sqrt{\mu^2+16(k-1)^2\sigma^2}}{2}, \infty\right)$. By Lemma 2.8.4, for $r > (2k-1)^2$, and for $Tx = (x, x^2, \cdots, x^{2k-1})$

$$\begin{aligned}
&\left\| \int_{\mathbb{R}} e^{i\zeta^T Tx} f_{\mathcal{N}}(x|\mu,\sigma) dx \right\|_{L^r} \\
\leq & C(r) \left(\frac{|\mu| + \sqrt{\mu^2 + 16(k-1)^2\sigma^2}}{2} + 2 \right)^{4(k-1)^2} \\
& \left(\| |x|^{4(k-1)^2} f_{\mathcal{N}}(x|\mu,\sigma) \|_{L^1} + \frac{1}{\sqrt{2\pi}\sigma} + \left\| (|x|^{4(k-1)^2} + 1) \frac{\partial f_{\mathcal{N}}(x|\mu,\sigma)}{\partial x} \right\|_{L^1} + 1 \right) \\
:= & h_3(\mu,\sigma),
\end{aligned}$$

where $C(r)$ is a constant that depends only r. It can be verified easily by the dominated convergence theorem that $h_3(\mu,\sigma)$ is a continuous function of μ. Then

$$\|\phi_T(\zeta|\theta)\|_{L^r} \leq \sum_{i=1}^{k} \pi_i \left\| \int_{\mathbb{R}} e^{i\zeta^T Tx} f_{\mathcal{N}}(x|\mu_i,\sigma) dx \right\|_{L^r} \leq \sum_{i=1}^{k} \pi_i h_3(\mu_i,\sigma),$$

which is a finite continuous function of $\theta = (\pi_1, \ldots, \pi_{k-1}, \mu_1, \ldots, \mu_k)$. Thus (A3) in Definition 2.5.13 for T is verified.

In conclusion, we have verified that T is admissible with respect to Θ. In the next subsection we continue to verify that the mean map $\lambda(\theta)$ is injective and its Jacobian is of full column rank.

2.8.3 Moment map is injective and its Jacobian is of full column rank

We first show that $\lambda(\theta)$ is injective. By (2.41), for any $j \in [2k-1]$:

$$\lambda^{(j)}(\theta) = \sum_{i=1}^{k} \pi_i(\mu_i^j + \sum_{\ell=1}^{\ell} \sigma^\ell \mathbb{E}Y^\ell \mu_i^{j-\ell}) = \sum_{i=1}^{k} \pi_i(\mu_i^j + \sum_{\substack{\ell=2 \\ \ell \text{ even}}}^{j} \sigma^\ell (\ell-1)!! \mu_i^{j-\ell})$$
$$= \sum_{i=1}^{k} \pi_i \mu_i^j + \sum_{\substack{\ell=2 \\ \ell \text{ even}}}^{j} \sigma^\ell (\ell-1)!! \sum_{i=1}^{k} \pi_i \mu_i^{j-\ell}. \quad (2.49)$$

Thus $\lambda(\theta) = \lambda(\bar{\theta})$ if and only if

$$\sum_{i=1}^{k} \pi_i \mu_i^j = \sum_{i=1}^{k} \bar{\pi}_i \bar{\mu}_i^j, \quad \forall j \in [2k-1] \cup \{0\}, \quad (2.50)$$

where the equation for $j = 0$ is due to $\sum_{i=1}^{k} \pi_i = \sum_{i=1}^{k} \bar{\pi}_i = 1$. Suppose that $\tilde{k} = |\{\mu_i\}_{i=1}^k \cap \{\bar{\mu}_i\}_{i=1}^k| < k$, and let $\{\mu_i\}_{i=1}^k \cup \{\bar{\mu}_i\}_{i=1}^k = \{\tilde{\mu}_i\}_{i=1}^{2k-\tilde{k}}$ be the distinct elements. Then the preceding displays can be written as

$$\sum_{i=1}^{\tilde{k}} a_i \tilde{\mu}_i^j = 0, \quad \forall j \in [k] \cup \{0\}. \quad (2.51)$$

There exists some $\tilde{\mu}_i \notin \{\mu_i\}_{i=1}^k \cap \{\bar{\mu}_i\}_{i=1}^k$ and its coefficient a_i is either π_ℓ or $-\bar{\pi}_\ell$ for some $\ell \in [k]$. But the coefficients of the first $\tilde{k}$ equations ($j = 0, 1, \ldots, \tilde{k}-1$) of system of linear equations (2.51) form a Vandermonde matrix and thus all $a_i = 0$ for $i \in [\tilde{k}]$. That means for some $\ell \in [k]$, $\pi_\ell = 0$ or $\bar{\pi}_\ell = 0$, which is a contradiction. Hence $\tilde{k} = k$ and consequently $\mu_i = \bar{\mu}_i$ for $i \in [k]$. Then system of equations (2.50) become

$$\sum_{i=1}^{k} (\pi_i - \bar{\pi}_i) \mu_i^j = 0, \quad \forall j \in [2k-1] \cup \{0\}.$$

Since the coefficients of the first k equations ($j = 0, 1, \ldots, k-1$) form a Vandermonde matrix, the unique solution is $\pi_i = \bar{\pi}_i$ for $i \in [k]$. In conclusion, $\theta = \bar{\theta}$, which establishes that $\lambda(\theta)$ is injective.

Next we show that $J_\lambda(\theta)$ is of full column rank for any $\theta \in \Theta$, where $J_\lambda(\theta)$ is the Jacobian matrix of $\lambda(\theta)$. Denote $\bar{\lambda}^{(j)}(\theta) = \sum_{i=1}^{k} \pi_i \mu_i^j$ and $\bar{\lambda}(\theta) = (\bar{\lambda}^{(1)}(\theta), \ldots, \bar{\lambda}^{(2k-1)}(\theta)) \in \mathbb{R}^{2k-1}$. By (2.49), $\lambda^{(j)}(\theta) = \bar{\lambda}^{(j)}(\theta) + \sum_{\substack{\ell=2 \\ \ell \text{ even}}}^{j} \sigma^\ell (\ell-1)!! \bar{\lambda}^{(j-\ell)}(\theta)$, which implies

$$\nabla_\theta \lambda^{(j)}(\theta) = \nabla_\theta \bar{\lambda}^{(j)}(\theta) + \sum_{\substack{\ell=2 \\ \ell \text{ even}}}^{j} \sigma^\ell (\ell-1)!! \nabla_\theta \bar{\lambda}^{(j-\ell)}(\theta).$$

Since $\nabla_\theta \lambda^{(j)}(\theta)$ and $\nabla_\theta \bar{\lambda}^{(j)}(\theta)$ are respectively the j-th row of $J_\lambda(\theta)$ and $J_{\bar{\lambda}}(\theta)$,

$$\det(J_\lambda(\theta)) = \det(J_{\bar{\lambda}}(\theta)). \tag{2.52}$$

Also, observe

$$\begin{aligned}
&\det(J_{\bar{\lambda}}(\theta)) \\
&= \left(\prod_{\ell=1}^{k} \pi_\ell\right) \det \begin{pmatrix} \mu_1 - \mu_k, & \dots & \mu_{k-1} - \mu_k, & 1, & \dots & 1 \\ \mu_1^2 - \mu_k^2, & \dots & \mu_{k-1}^2 - \mu_k^2, & 2\mu_1, & \dots & 2\mu_k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_1^{2k-1} - \mu_k^{2k-1}, & \dots & \mu_{k-1}^{2k-1} - \mu_k^{2k-1}, & (2k-1)\mu_1^{2k-1}, & \dots & (2k-1)\mu_k^{2k-1} \end{pmatrix} \\
&= \left(\prod_{\ell=1}^{k} \pi_\ell\right) (-1)^{k+1} \det \begin{pmatrix} 1, & \dots & 1, & 1, & 0, & \dots & 0 \\ \mu_1, & \dots & \mu_{k-1}, & \mu_k, & 1, & \dots & 1 \\ \mu_1^2, & \dots & \mu_{k-1}^2, & \mu_k^2 & 2\mu_1, & \dots & 2\mu_k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_1^{2k-1}, & \dots & \mu_{k-1}^{2k-1}, & \mu_k^{2k-1} & (2k-1)\mu_1^{2k-1}, & \dots & (2k-1)\mu_k^{2k-1} \end{pmatrix} \\
&= \left(\prod_{\ell=1}^{k} \pi_\ell\right) (-1)^{k+1} \left(\prod_{i=1}^{k} (-1)^{k+i-2i}\right) \prod_{1 \le \alpha < \beta \le k} (\mu_\alpha - \mu_\beta)^4
\end{aligned} \tag{2.53}$$

where the second equality holds since we may subtract the k-th column of the $2k \times 2k$ matrix from each of its first $k-1$ columns and then do Laplace expansion along its first row, and the last equality follows by observing that the $(k+i)$-th column of the $2k \times 2k$ matrix is the derivative of the i-th column and by Lemma 2.5.11 c) after some column permutation.

By (2.52) and (2.53), $\det(J_\lambda(\theta)) \neq 0$ on Θ. That is $J_\lambda(\theta)$ is of full column rank for any $\theta \in \Theta$.

In summary, we have showed that $\lambda(\theta)$ is injective and its Jacobian is of full column rank, which means the condition 1) and 2) in Corollary 2.5.15 are satisfied. Together with the preceding subsection, all the conditions in Corollary 2.5.15 and by Corollary 2.5.15, for P_θ having the density in (2.40), the inverse bounds (2.22) and (2.24) hold for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$.

2.8.4 Kernel P_θ is mixture of Dirichlet processes

Using the tools developed in subsection 2.8.1, we are now in a proper position to complete Example 2.5.21, which is motivated from modeling techniques in nonparametric Bayesian statistics. In particular, the kernel P_θ is now defined as a distribution on a space of probability measures: P_θ is a mixture of Dirichlet processes, so that $P_{G,N}$ is in fact a finite mixture of products of mixtures of Dirichlet processes.

Let $\mathfrak{X} = \mathscr{P}(\mathfrak{Z})$ be the space of all probability measures on a Polish space $(\mathfrak{Z}, \mathscr{Z})$. $\mathfrak{X}$ is equipped with the weak topology and the corresponding Borel sigma algebra $\mathcal{A}$. Let $\mathscr{D}_{\alpha H}$ denote the Dirichlet distribution on $(\mathfrak{X}, \mathcal{A})$, which is specified by two parameters, concentration parameter $\alpha \in (0, \infty)$ and base measure $H \in \mathfrak{X}$. Formal definition and key properties of the Dirichlet distributions can be found in the original chapter of [Fer73], or a recent textbook by [GvdV17]. In this example, we take the probability kernel P_θ to be a mixture of two Dirichlet distributions with different concentration parameters, while the base measure is fixed and known: $P_\theta = \pi_1 \mathscr{D}_{\alpha_1 H} + (1 - \pi_1)\mathscr{D}_{\alpha_2 H}$. Thus, the parameter vector is three dimensional which shall be restricted by the following constraint: $\theta := (\pi_1, \alpha_1, \alpha_2) \in \Theta = \{(\pi_1, \alpha_1, \alpha_2) | 0 < \pi_1 < 1, 2 < \alpha_1 < \alpha_2\}$. Kernel P_θ so defined is a simple instance of the so-called mixture of Dirichlet processes first studied by [Ant74], but considerably more complex instances of model using Dirichlet as the building block have become a main staple in the lively literature of Bayesian nonparametrics [HHMW10, TJBB06, RDG08, CLOP19]. For notational convenience in the following we also denote $Q_\alpha := \mathscr{D}_{\alpha H}$ for $\alpha = \alpha_1$ and $\alpha = \alpha_2$, noting that H is fixed, so we may write $P_\theta = \pi_1 Q_{\alpha_1} + (1 - \pi_1) Q_{\alpha_2}$.

Having specified the kernel P_θ, now let $G \in \mathcal{E}_k(\Theta)$. The mixture of product distributions $P_{G,N}$ is defined in the same way as before (cf. Eq. (2.1)). Now we shall show that for $G_0 \in \mathcal{E}_{k_0}(\Theta^\circ) = \mathcal{E}_{k_0}(\Theta)$, (2.22) and (2.24) hold by an application of Corollary 2.5.15 via a suitable choice of map T.

Consider a map $T : \mathfrak{X} \to \mathbb{R}^3$ defined by $Tx = ((x(B))^2, (x(B))^3, (x(B))^4)^T$ for some $B \in \mathscr{Z}$ to be specified later. The reason we restrict the domain of Θ is so that this particular choice of map will be shown to be admissible. Define $T_1 : \mathfrak{X} \to \mathbb{R}$ by $T_1 x = x(B)$ and $T_2 : \mathbb{R} \to \mathbb{R}^3$ by $T_2 z = (z^2, z^3, z^4)^T$. Then $T = T_2 \circ T_1$. For $X \sim P_\theta$, $T_1 X$ has induced probability measure on $\mathbb{R}$

$$P_\theta \circ T_1^{-1} = \pi_1 \left(Q_{\alpha_1} \circ T_1^{-1}\right) + \pi_2 \left(Q_{\alpha_2} \circ T_1^{-1}\right).$$

where $\pi_2 = 1 - \pi_1$. By a standard property of Dirichlet distribution, as $Q_\alpha = \mathscr{D}_{\alpha H}$, we have $Q_\alpha \circ T_1^{-1}$ corresponds to Beta$(\alpha H(B), \alpha(1 - H(B)))$, a Beta distribution with parameter induced from the Dirichlet distribution. Thus with $\xi = H(B)$, $Q_\alpha \circ T_1^{-1}$ has density w.r.t. Lebesgue measure on $\mathbb{R}$

$$g(z|\alpha, \xi) = \frac{1}{B(\alpha\xi, \alpha(1-\xi))} z^{\alpha\xi - 1}(1 - z)^{\alpha(1-\xi)-1} \mathbf{1}_{(0,1)}(z),$$

where $B(\cdot, \cdot)$ is the beta function. Then $P_\theta \circ T_1^{-1}$ has density w.r.t. Lebesgue measure $\pi_1 g(z|\alpha_1, \xi) + \pi_2 g(z|\alpha_2, \xi)$.

Now, the push-forward measure $P_\theta \circ T^{-1} = (P_\theta \circ T_1^{-1}) \circ T_2^{-1}$ has mean $\lambda(\theta) \in \mathbb{R}^3$ with

$$\lambda^{(j)}(\theta) = \sum_{i=1}^{2} \pi_i \int_{\mathbb{R}} z^{j+1} g(z|\alpha_i, \xi) dz = \sum_{i=1}^{2} \pi_i \prod_{\ell=0}^{j} \frac{\alpha_i \xi + \ell}{\alpha_i + \ell} \quad \forall j = 1, 2, 3$$

and has covariance matrix Λ with its $j\beta$ entry given by

$$\Lambda_{j\beta}(\theta) = \sum_{i=1}^{2} \pi_i \int_{\mathbb{R}} z^{j+\beta+2} g(z|\alpha_i, \xi) dz - \lambda^j(\theta)\lambda^\beta(\theta) = \sum_{i=1}^{2} \pi_i \prod_{\ell=0}^{j+\beta+1} \frac{\alpha_i \xi + \ell}{\alpha_i + \ell} - \lambda^j(\theta)\lambda^\beta(\theta).$$

It follows immediately from these formula that $\lambda(\theta)$ and $\Lambda(\theta)$ are continuous on Θ. That is, (A1) in Definition 2.5.13 is satisfied. Furthermore, observe that $P_\theta \circ T^{-1}$ has characteristic function

$$\phi_T(\zeta|\theta) = \pi_1 h(\zeta|\alpha_1, \xi) + \pi_2 h(\zeta|\alpha_2, \xi)$$

where $h(\zeta|\alpha, \xi) = \int_{\mathbb{R}} \exp(\boldsymbol{i} \sum_{j=1}^{3} \zeta^{(j)} z^j) g(z|\alpha, \xi) dz$. The verification of (A2) in Definition 2.5.13 is omitted since the essence is the same as the next four equations due to the dominated convergence theorem.

It is easy to verify by the dominated convergence theorem or Pratt's Lemma:

$$\begin{aligned}
\frac{\partial h(\zeta|\alpha, \xi)}{\partial \alpha} &= \int_{\mathbb{R}} \exp(\boldsymbol{i} \sum_{i=1}^{3} \zeta^{(i)} z^{i+1}) \frac{\partial g(z|\alpha, \xi)}{\partial \alpha} dz, \\
\frac{\partial^2 h(\zeta|\alpha, \xi)}{\partial \alpha^2} &= \int_{\mathbb{R}} \exp(\boldsymbol{i} \sum_{i=1}^{3} \zeta^{(i)} z^{i+1}) \frac{\partial^2 g(z|\alpha, \xi)}{\partial \alpha^2} dz, \\
\frac{\partial h(\zeta|\alpha, \xi)}{\partial \zeta^{(j)}} &= \int_{\mathbb{R}} \boldsymbol{i} z^{j+1} \exp(\boldsymbol{i} \sum_{i=1}^{3} \zeta^{(i)} z^{i+1}) g(z|\alpha, \xi) dz, \quad j = 1, 2, 3
\end{aligned}$$

and

$$\frac{\partial^2 h(\zeta|\alpha, \xi)}{\partial \zeta^{(j)} \partial \alpha} = \frac{\partial^2 h(\zeta|\alpha, \xi)}{\partial \alpha \partial \zeta^{(j)}} = \int_{\mathbb{R}} \boldsymbol{i} z^{j+1} \exp(\boldsymbol{i} \sum_{i=1}^{3} \zeta^{(i)} z^{i+1}) \frac{\partial g(z|\alpha, \xi)}{\partial \alpha} dz, \quad j = 1, 2, 3.$$

From the preceding four displays,

$$\left| \frac{\partial h(\zeta|\alpha, \xi)}{\partial \alpha} \right| \leq \int_{\mathbb{R}} \left| \frac{\partial g(z|\alpha, \xi)}{\partial \alpha} \right| dz := h_1(\alpha) \tag{2.54}$$

$$\left| \frac{\partial^2 h(\zeta|\alpha, \xi)}{\partial \alpha^2} \right| \leq \int_{\mathbb{R}} \left| \frac{\partial^2 g(z|\alpha, \xi)}{\partial \alpha^2} \right| dz := h_2(\alpha), \tag{2.55}$$

$$\max_{j=1,2,3} \left| \frac{\partial h(\zeta|\alpha, \xi)}{\partial \zeta^{(j)}} \right| \leq \max_{j=1,2,3} \int_{\mathbb{R}} \left| z^{j+1} g(z|\alpha, \xi) \right| dz := h_3(\alpha), \tag{2.56}$$

$$\max_{j=1,2,3} \left| \frac{\partial^2 h(\zeta|\alpha, \xi)}{\partial \zeta^{(j)} \partial \alpha} \right| \leq \max_{j=1,2,3} \int_{\mathbb{R}} \left| z^{j+1} \frac{\partial g(z|\alpha, \xi)}{\partial \alpha} \right| dz := h_4(\alpha), \tag{2.57}$$

where $h_1(\alpha)$, $h_2(\alpha)$, $h_3(\alpha)$ and $h_4(\alpha)$ are continuous functions of α by dominated convergence

theorem, with their dependence on the constant ξ suppressed.

It follows that gradient w.r.t. θ is

$$\nabla_\theta \phi_T(\zeta|\theta) = \left(h(\zeta|\alpha_1,\xi) - h(\zeta|\alpha_2,\xi), \pi_1 \frac{\partial h(\zeta|\alpha_1,\xi)}{\partial \alpha}, \pi_2 \frac{\partial h(\zeta|\alpha_2,\xi)}{\partial \alpha} \right)^T, \tag{2.58}$$

and Hessian w.r.t. θ is

$$\mathbf{Hess}_\theta \phi_T(\zeta|\theta) = \begin{pmatrix} 0 & \frac{\partial h(\zeta|\alpha_1,\xi)}{\partial \alpha} & -\frac{\partial h(\zeta|\alpha_2,\xi)}{\partial \alpha} \\ \frac{\partial h(\zeta|\alpha_1,\xi)}{\partial \alpha} & \pi_1 \frac{\partial^2 h(\zeta|\alpha_1,\xi)}{\partial \alpha^2} & 0 \\ -\frac{\partial h(\zeta|\alpha_2,\xi)}{\partial \alpha} & 0 & \pi_2 \frac{\partial^2 h(\zeta|\alpha_2,\xi)}{\partial \alpha^2} \end{pmatrix}. \tag{2.59}$$

Then by (2.54), (2.55), (2.56), (2.57), (2.58) and (2.59), for any $i, j \in \{1,2,3\}$:

$$\begin{aligned} \left| \frac{\partial \phi_T(\zeta|\theta)}{\partial \theta^{(i)}} \right| &\leq 2 + h_1(\alpha_1) + h_1(\alpha_2), \\ \left| \frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \theta^{(i)} \partial \theta^{(j)}} \right| &\leq \sum_{i=1}^{2} \left(h_1(\alpha_i) + h_2(\alpha_i) \right), \\ \left| \frac{\partial^2 \phi_T(\zeta|\theta)}{\partial \zeta^{(j)} \partial \theta^{(i)}} \right| &\leq \sum_{i=1}^{2} \left(h_3(\alpha_i) + h_4(\alpha_i) \right), \end{aligned}$$

where the right hand side of the preceding 3 displays are continuous functions of θ since h_1, h_2, h_3 and h_4 are continuous.

So far we have shown some properties of T for every B. For some other properties we will need to specify B. For $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ with $\theta_i^0 = (\pi_{1i}^0, \alpha_{1i}^0, \alpha_{2i}^0) \in \Theta$, let B be such that $\xi = H(B) \in (1/\min_{i\in[k_0]} \alpha_{1i}^0, 1 - 1/\min_{i\in[k_0]} \alpha_{1i}^0)$. Notice that since $\alpha_{1i}^0 > 2$, $(1/\min_{i\in[k_0]} \alpha_{1i}^0, 1 - 1/\min_{i\in[k_0]} \alpha_{1i}^0)$ is not empty. Hence to verify the condition (A3) in Definition 2.5.13 w.r.t. $\{\theta_i^0\}_{i=1}^{k_0}$ for T with the B specified it suffices to establish there exists some $r \geq 1$ such that $\int_{\mathbb{R}^3} |\phi_T(\zeta|\theta)|^r \, d\zeta$ in a small neighborhood of θ_0 is upper bounded by a finite continuous function of θ for each $\theta_0 \in \{\theta_i^0\}_{i=1}^{k_0}$.

Since $g(z|\alpha,\xi)$ is differentiable w.r.t. to z on $\mathbb{R}\backslash\{0,1\}$ and when $z \neq 0, 1$

$$\begin{aligned} &\frac{\partial g(z|\alpha,\xi)}{\partial z} \\ =&\frac{\mathbf{1}_{(0,1)}(z)}{B(\alpha\xi, \alpha(1-\xi))} \left((\alpha\xi - 1) z^{\alpha\xi - 2} (1-z)^{\alpha(1-\xi)-1} - (\alpha(1-\xi) - 1) z^{\alpha\xi - 1} (1-z)^{\alpha(1-\xi)-2} \right), \end{aligned}$$

which is in L^1 when $\alpha \geq \min_{i\in[k_0]} \alpha_{1i}^0 - \gamma$ such that $\alpha\xi > 1$ and $\alpha(1-\xi) > 1$, where γ depends on T through ξ. Moreover, $g(z|\alpha,\xi)$ and $z^2 g(z|\alpha,\xi)$ are both increasing on $(-\infty, -1)$ and decreasing

on $(1,\infty)$. Now, by appealing to Lemma 2.8.4, for $r>12$, and for $\alpha \geq \min_{i\in[k_0]}\alpha_{1i}^0-\gamma$

$$\begin{aligned}
&\|h(\zeta|\alpha,\xi)\|_{L^r}\\
\leq& C(r)(1+2)^2\left(\|z^2 g(z|\alpha,\xi)\|_{L^1}+3\|g(z|\alpha,\xi)\|_{L^\infty}+\left\|(z^2+1)\frac{\partial g(z|\alpha,\xi)}{\partial z}\right\|_{L^1}+1\right)\\
:=& h_5(\alpha,\xi),
\end{aligned}$$

where $C(r)$ is a constant that depends only on r. It can be verified easily by the dominated convergence theorem that $h_5(\alpha,\xi)$ is a continuous function of α. Then for θ in a neighborhood of $\theta_0\in\{\theta_i^0\}_{i=1}^{k_0}$ such that $\alpha_1,\alpha_2\geq\alpha_{1i}^0-\gamma$,

$$\begin{aligned}
&\|\phi_T(\zeta|\theta)\|_{L^r}\\
\leq&\pi_1\|h(\zeta|\alpha_1,\xi)\|_{L^r}+\pi_2\|h(\zeta|\alpha_2,\xi)\|_{L^r}\\
\leq&\pi_1 h_5(\alpha_1,\xi)+\pi_2 h_5(\alpha_2,\xi),
\end{aligned}$$

which is a finite continuous function of $\theta=(\pi_1,\alpha_1,\alpha_2)$. We have thus verified that T with the specified B is admissible w.r.t. $\{\theta_i^0\}_{i=1}^{k_0}$.

Moreover, it can also be verified that $\lambda(\theta)$ for T is injective on Θ provided that $\xi\neq\frac{1}{3},\frac{1}{2},\frac{2}{3}$. By calculation,

$$\det(J_\lambda)(\theta)=-\frac{6(\xi-1)^3\xi^3(2\xi-1)(3\xi-1)(3\xi-2)\pi_1\pi_2(\alpha_1-\alpha_2)^4}{\prod_{i=1}^2((1+\alpha_i)^2(2+\alpha_i)^2(3+\alpha_i)^2)}\neq 0$$

on Θ provided that $\xi\neq\frac{1}{3},\frac{1}{2},\frac{2}{3}$; so $J_\lambda(\theta)$ is of full rank for each $\theta\in\Theta$ provided that $\xi\neq\frac{1}{3},\frac{1}{2},\frac{2}{3}$. In summary, for $G_0=\sum_{i=1}^{k_0}p_i^0\delta_{\theta_i^0}$ with $\theta_i^0=(\pi_{1i}^0,\alpha_{1i}^0,\alpha_{2i}^0)\in\Theta$, $Tx=((x(B))^2,(x(B))^3,(x(B))^4)^T$ with B such that $\xi=H(B)\in(\frac{1}{\min_{i\in[k_0]}\alpha_{1i}^0},1-\frac{1}{\min_{i\in[k_0]}\alpha_{1i}^0})\backslash\{\frac{1}{3},\frac{1}{2},\frac{2}{3}\}$ satisfies all the conditions in Corollary 2.5.15 and thus (2.22) and (2.24) hold.

2.9 Proofs of lemmas in Section 2.3

Proof of Lemma 2.3.2: a) The proof is trivial and is therefore omitted.

b) Let $G=\sum_{i=1}^k p_i\delta_{\theta_i}$ and $G'=\sum_{i=1}^k p_i'\delta_{\theta_i'}$ be their increasing representation. Let τ be the optimal permutation that achieves $D_1(G,G')=\sum_{i=1}^k\left(\|\theta_{\tau(i)}-\theta_i'\|_2+|p_{\tau(i)}-p_i'|\right)$. Let $\boldsymbol{q}$ be a coupling of the mixing probabilities $\mathbf{p}=(p_1,\ldots,p_k)$ and $\mathbf{p}'=(p_1',\ldots,p_k')$ such that $q_{\tau(i),i}=\min\{p_{\tau(i)},p_i\}$ and then the remaining mass to be assigned is $\sum_{i=1}^k(p_{\tau(i)}-q_{\tau(i),i})=$

$\frac{1}{2}\sum_{i=1}^{k}|p_{\tau(i)}-p_i|$. Thus,

$$\begin{aligned}&W_1(G,G')\\&\leq\sum_{i=1}^{k}q_{\tau(i),i}\|\theta_{\tau(i)}-\theta'_i\|_2+\frac{1}{2}\sum_{i=1}^{k}|p_{\tau(i)}-p_i|\mathrm{diam}(\Theta)\leq\max\left\{1,\frac{\mathrm{diam}(\Theta)}{2}\right\}D_1(G,G').\end{aligned}$$

The proof for the general case proceed in the same procedure.

c) Let $\theta\in\Theta^\circ$ and then for sufficiently small positive a, $\theta+ae_1\in\Theta$ where $e_1=(1,0,\ldots,0)$ is the first canonical basis. Consider $G=\sum_{i=1}^{k}p_i\delta_{\theta_i}$ with $p_i=\frac{1}{k}$ for $1\leq i\leq k$ and $\theta_{k-1}=\theta$ and $\theta_k=\theta+ae_1$, and $\{\theta_i\}_{i=1}^{k-2}$ be any arbitrary distinct $k-2$ points that are different from θ and $\theta+ae_1$. Let $G'=\sum_{i=1}^{k}p'_i\delta_{\theta'_i}$ with $p'_i=p_i$ for $1\leq i\leq k-2$, $p'_{k-1}=\frac{1}{k}+\frac{a}{2}\|e_1\|_2$, $p'_k=\frac{1}{k}-\frac{a}{2}\|e_1\|_2$ and $\theta'_i=\theta_i$ for $1\leq i\leq k$.

Let $\boldsymbol{q}$ be a coupling of $\boldsymbol{p}=(p_1,\ldots,p_k)$ and $\boldsymbol{p}'=(p'_1,\ldots,p'_{k-2},p'_{k-1},p'_k)$ such that

$$q_{ij}=\begin{cases}\frac{1}{k} & 1\leq i=j\leq k-1,\\ \frac{1}{k}-\frac{a}{2}\|e_1\|_2 & i=j=k,\\ \frac{a}{2}\|e_1\|_2 & i=k,j=k-1,\\ 0 & \text{otherwise.}\end{cases}$$

Then

$$W_1(G,G')\leq\sum_{i,j}q_{ij}\|\theta_i-\theta'_j\|_2=\frac{a}{2}\|e_1\|_2\times a\|e_1\|_2.$$

Moreover, it's easy to see $D_1(G,G')=a\|e_1\|_2$ when a is sufficiently small. Thus

$$\inf_{G,G'\in\mathcal{E}_k(\Theta)}\frac{W_1(G,G')}{D_1(G,G')}\leq\lim_{a\to0}\frac{a}{2}\|e_1\|_2=0.$$

d) Consider any $G_n\in\mathcal{E}_k(\Theta)$ and $G_n\overset{W_1}{\to}G_0$, and one may write $G_n=\sum_{i=1}^{k}p_i^n\delta_{\theta_i^n}$ for $n\geq0$ such that $p_i^n\to p_i^0$ and $\theta_i^n\to\theta_i^0$. Then when n is sufficiently large, $G_n\in\mathcal{E}_k(\Theta_1)$ for $\Theta_1=\bigcup_{i=1}^{k_0}B(\theta_i^0,\frac{1}{2})$, where $B(\theta_i^0,\rho)\subset\mathbb{R}^q$ is the open ball with center at θ_i^0 of radius ρ. Then by b) for large n, $W_1(G_n,G_0)\leq C(G_0)D_1(G_n,G_0)$, which entails $\liminf_{\substack{G\overset{W_1}{\to}G_0\\G\in\mathcal{E}_k(\Theta)}}\frac{D_1(G,G_0)}{W_1(G,G_0)}>0$.

On the other hand, for sufficiently large n, one can verify

$$W_1(G_n,G_0)\geq\frac{1}{2}\sum_{j=1}^{k}p_j^0\|\theta_j^n-\theta_j^0\|_2+\frac{1}{8}\min_{1\leq i<\ell\leq k}\|\theta_i^0-\theta_\ell^0\|_2\sum_{j=1}^{k}|p_j^n-p_j^0|$$

$$\geq \frac{1}{2}\min\left\{\min_{\ell} p_\ell^0, \frac{1}{4}\min_{1\leq i<\ell\leq k}\|\theta_i^0-\theta_\ell^0\|_2\right\}\sum_{j=1}^{k}(\|\theta_j^n-\theta_j^0\|_2+|p_j^n-p_j^0|)$$
$$= \frac{1}{2}\min\left\{\min_{\ell} p_\ell^0, \frac{1}{4}\min_{1\leq i<\ell\leq k}\|\theta_i^0-\theta_\ell^0\|_2\right\} D_1(G_n, G_0),$$

which entails $\liminf_{\substack{G\xrightarrow{W_1}G_0\\ G\in\mathcal{E}_k(\Theta)}} \frac{W_1(G,G_0)}{D_1(G,G_0)} > 0$.

e) Based on d), there exists $c(G_0) > 0$ such that for $G \in \mathcal{E}_{k_0}(\Theta)$ satisfying $W_1(G, G_0) < c(G_0)$

$$W_1(G, G_0) \geq C_1(G_0) D_1(G, G_0).$$

For $G \in \mathcal{E}_{k_0}(\Theta)$ satisfying $W_1(G, G_0) \geq c(G_0)$

$$\frac{W_1(G, G_0)}{D_1(G, G_0)} \geq \frac{c(G_0)}{k_0 \text{diam}(\Theta) + 1}.$$

□

2.10 Proofs and auxiliary lemmas of Section 2.4

2.10.1 Additional examples and proofs of results in Section 2.4.1

Example 2.10.1 (Location Gamma kernel)**.** For gamma distribution with fixed $\alpha \in (0,1)\bigcup(1,2)$ and $\beta > 0$, consider its location family with density $f(x|\theta) = \frac{\beta^\alpha (x-\theta)^{\alpha-1} e^{-\beta(x-\theta)}}{\Gamma(\alpha)} \mathbf{1}_{(\theta,\infty)}(x)$ w.r.t. Lebesgue measure μ on $\mathfrak{X} = \mathbb{R}$. The parameter space $\Theta = \mathbb{R}$. Observe

$$\lim_{a\to 0^+} \frac{f(\theta_0|\theta_0+a) - f(\theta_0|\theta_0)}{a} = 0$$

and

$$\lim_{a\to 0^+} \frac{f(\theta_0|\theta_0-a) - f(\theta_0|\theta_0)}{a} = \frac{\beta^\alpha}{\Gamma(\alpha)} \lim_{a\to 0^+} a^{\alpha-2} e^{-\beta a} = \infty,$$

since $\alpha < 2$. Then for any x, $f(x|\theta)$ as a function of θ is not differentiable at $\theta = x$. So it's not identifiable in the first order as in [HN16b].

However, this family does satisfy the $(\{\theta_i\}_{i=1}^k, \mathcal{N})$ first-order identifiable definition with $\mathcal{N} = \bigcup_{i=1}^k (\theta_i - \rho, \theta_i + \rho)$ where $\rho = \frac{1}{4}\min_{1\leq i<j\leq k}|\theta_i - \theta_j|$. Indeed, observing

$$\frac{\partial}{\partial\theta} f(x|\theta) = \left(\beta - \frac{\alpha-1}{x-\theta}\right) f(x|\theta), \quad \forall \theta \neq x,$$

then (2.9a) become

$$0 = \sum_{i=1}^{k} \left(a_i \frac{\partial}{\partial \theta} f(x|\theta_i) + b_i f(x|\theta_i) \right) = \sum_{i=1}^{k} \left(a_i \beta - a_i \frac{\alpha - 1}{x - \theta_i} + b_i \right) f(x|\theta_i) \quad \text{for } \mu\text{−}a.e.\, x \in \mathbb{R} \backslash \mathcal{N}.$$

Without loss of generality, assume $\theta_1 < \ldots < \theta_k$. Then for $\mu\text{−}a.e.\, x \in (\theta_1, \theta_2) \backslash \mathcal{N} = [\theta_1 + \rho, \theta_2 - \rho]$, the above display become

$$\left(a_1 \beta - a_1 \frac{\alpha - 1}{x - \theta_1} + b_1 \right) \frac{\beta^\alpha (x - \theta_1)^{\alpha - 1} e^{-\beta(x - \theta_1)}}{\Gamma(\alpha)} = 0$$

which implies $a_1 = b_1 = 0$ since $\alpha \neq 1$. Repeating the above argument on interval $(\theta_2, \theta_3), \ldots (\theta_k, \infty)$ shows $a_i = b_i = 0$ for any $i \in [k]$.

So for $\{\theta_i\}_{i=1}^k$ this family is $(\{\theta_i\}_{i=1}^k, \mathcal{N})$ first-order identifiable. Moreover, for every x in $\mathbb{R} \backslash \mathcal{N}$ $f(x|\theta)$ is continuously differentiable w.r.t. θ in a neighborhood of θ_i^0 for $i \in [k_0]$. By Lemma 2.4.2 b) for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$ (2.12) holds. □

Proof of Lemma 2.4.2 b): Suppose the equation (2.12) is incorrect. Then there exists $G_\ell, H_\ell \in \mathcal{E}_{k_0}(\Theta)$ such that

$$\begin{cases} G_\ell \neq H_\ell, & \forall \ell \\ G_\ell, H_\ell \overset{W_1}{\to} G_0, & \text{as } \ell \to \infty \\ \frac{V(P_{G_\ell}, P_{H_\ell})}{D_1(G_\ell, H_\ell)} \to 0, & \text{as } \ell \to \infty. \end{cases}$$

We may relabel the atoms of G_ℓ and H_ℓ such that $G_\ell = \sum_{i=1}^{k_0} p_i^\ell \delta_{\theta_i^\ell}$, $H_\ell = \sum_{i=1}^{k_0} \pi_i^\ell \delta_{\eta_i^\ell}$ with $\theta_i^\ell, \eta_i^\ell \to \theta_i^0$ and $p_i^\ell, \pi_i^\ell \to p_i^0$ as $\ell \to \infty$ for any $i \in [k_0]$. With subsequences argument if necessary, we may futher require

$$\frac{\theta_i^\ell - \eta_i^\ell}{D_1(G_\ell, H_\ell)} \to a_i \in \mathbb{R}^q, \quad \frac{p_i^\ell - \pi_i^\ell}{D_1(G_\ell, H_\ell)} \to b_i \in \mathbb{R}, \quad \forall 1 \leq i \leq k_0, \tag{2.60}$$

where b_i and the components of a_i are in $[-1, 1]$ and $\sum_{i=1}^{k_0} b_i = 0$. Moreover, $D_1(G_\ell, H_\ell) = \sum_{i=1}^{k_0} \left(\|\theta_i^\ell - \eta_i^\ell\|_2 + |p_i^\ell - \pi_i^\ell| \right)$ for sufficiently large ℓ, which implies

$$\sum_{i=1}^{k_0} \|a_i\|_2 + \sum_{i=1}^{k_0} |b_i| = 1.$$

It also follows that at least one of a_i is not $\mathbf{0} \in \mathbb{R}^q$ or one of b_i is not 0. On the other hand,

$$0 = \lim_{\ell \to \infty} \frac{2V(P_{G_\ell}, P_{G_0})}{D_1(G_\ell, G_0)}$$

$$
\begin{aligned}
&\geq \lim_{\ell\to\infty}\int_{\mathfrak{X}\setminus\mathcal{N}}\left|\sum_{i=1}^{k_0} p_i^\ell \frac{f(x|\theta_i^\ell)-f(x|\eta_i^\ell)}{D_1(G_\ell,H_\ell)}+\sum_{i=1}^{k_0} f(x|\eta_i^\ell)\frac{p_i^\ell-\pi_i^\ell}{D_1(G_\ell,H_\ell)}\right|\mu(dx)\\
&\geq \int_{\mathfrak{X}^d\setminus\mathcal{N}}\liminf_{\ell\to\infty}\left|\sum_{i=1}^{k_0} p_i^\ell \frac{f(x|\theta_i^\ell)-f(x|\eta_i^\ell)}{D_1(G_\ell,H_\ell)}+\sum_{i=1}^{k_0} f(x|\eta_i^\ell)\frac{p_i^\ell-\pi_i^\ell}{D_1(G_\ell,H_\ell)}\right|\mu(dx)\\
&= \int_{\mathfrak{X}\setminus\mathcal{N}}\left|\sum_{i=1}^{k_0} p_i^0 a_i^T\nabla_\theta f(x|\theta_i^0)+\sum_{i=1}^{k_0} f(x|\theta_i^0)b_i\right|\mu(dx),
\end{aligned}
$$

where the second inequality follows from Fatou's Lemma, and the last step follows from Lemma 2.10.3 a). Then $\sum_{i=1}^{k_0} p_i^0 a_i^T\nabla_\theta f(x|\theta_i^0)+\sum_{i=1}^{k_0} f(x|\theta_i^0)b_i=0$ for $\mu-a.e.\ x\in\mathfrak{X}\setminus\mathcal{N}$. Thus we find a nonzero solution to (2.9a), (2.9b) with k,θ_i replaced by k_0,θ_i^0.

However, the last statement contradicts with the definition of $(\{\theta_i^0\}_{i=1}^{k_0},\mathcal{N})$ first-order identifiable.

□

Proof of Lemma 2.4.4: By Lemma 2.4.14 b) $(a_1,b_1,\ldots,a_{k_0},b_{k_0})$ is also a nonzero solution of the system of equations (2.9a), (2.9b). Let $a_i'=\frac{a_i/p_i^0}{\sum_{i=1}^{k_0}\left(\|a_i/p_i^0\|_2+|b_i|\right)}$ and $b_i'=\frac{b_i}{\sum_{i=1}^{k_0}\left(\|a_i/p_i^0\|_2+|b_i|\right)}$. Then a_i' and b_i' satisfy $\sum_{i=1}^{k_0}\left(\|a_i'\|_2+|b_i'|\right)=1$ and $(p_1^0a_1',\ b_1',\ldots,p_{k_0}^0a_{k_0}',\ b_{k_0}')$ is also a nonzero solution of the system of equations (2.9a), (2.9b) with k,θ_i replaced respectively by k_0,θ_i^0. Let $G_\ell=p_i^\ell\delta_{\theta_i^\ell}$ with $p_i^\ell=p_i^0+b_i'\frac{1}{\ell}$ and $\theta_i^\ell=\theta_i^0+\frac{1}{\ell}a_i'$ for $1\leq i\leq k_0$. When ℓ is large, $0<p_i^\ell<1$ and $\theta_i^\ell\in\Theta$ since $0<p_i^0<1$ and $\theta_i^0\in\Theta^\circ$. Moreover, $\sum_{i=1}^{k_0}p_i^\ell=1$ since $\sum_{i=1}^{k_0}b_i'=0$. Then $G_\ell\in\mathcal{E}_{k_0}(\Theta)$ and $G_\ell\neq G_0$ since at least one of a_i' or b_i' is nonzero. When ℓ is large $D_1(G_\ell,G_0)=\sum_{i=1}^{k_0}\left(\|\theta_i^\ell-\theta_i^0\|_2+|p_i^\ell-p_i^0|\right)=\frac{1}{\ell}$. Thus when ℓ is large

$$
\frac{2V(P_{G_\ell},P_{G_0})}{D_1(G_\ell,G_0)}=\int_{\mathfrak{X}\setminus\mathcal{N}}\left|\sum_{i=1}^{k_0}p_i^\ell\frac{f(x|\theta_i^\ell)-f(x|\theta_i^0)}{1/\ell}+\sum_{i=1}^{k_0}b_i'f(x|\theta_i^0)\right|\mu(dx). \tag{2.61}
$$

Since by condition c)

$$
\left|\frac{f(x|\theta_i^\ell)-f(x|\theta_i^0)}{1/\ell}\right|=\left|\frac{f(x|\theta_i^0+\frac{1}{\ell}\frac{\|a_i'\|_2}{\|a_i\|_2}a_i)-f(x|\theta_i^0)}{1/\ell}\right|\leq\frac{\|a_i'\|_2}{\|a_i\|_2}\bar{f}(x),
$$

the integrand of (2.61) is bounded by $\sum_{i=1}^{k_0}\frac{1/p_i^0}{\sum_{i=1}^{k_0}\left(\|a_i/p_i^0\|_2+|b_i|\right)}\bar{f}(x)+\sum_{i=1}^{k_0}|b_i'|f(x|\theta_i^0)$, which is integrable w.r.t. to μ on $\mathfrak{X}\setminus\mathcal{N}$.

Then by the dominated convergence theorem

$$
\lim_{\ell\to\infty}\frac{2V(P_{G_\ell},P_{G_0})}{D_1(G_\ell,G_0)}=\int_{\mathfrak{X}\setminus\mathcal{N}}\left|\sum_{i=1}^{k_0}p_i^0\langle a_i',\nabla_\theta f(x|\theta_i^0)\rangle+\sum_{i=1}^{k_0}b_i'f(x|\theta_i^0)\right|\mu(dx)=0.
$$

Thus

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{D_1(G, G_0)} = 0.$$

and the proof is completed by

$$\lim_{r \to 0} \inf_{\substack{G,H \in B_{W_1}(G_0, r) \\ G \neq H}} \frac{V(P_G, P_H)}{D_1(G, H)} \leq \liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{D_1(G, G_0)}.$$

□

Proof of Lemma 2.4.7: The proof of b) is similar to that of a) and hence only the latter is presented. Let G_ℓ be the same as in the proof of Lemma 2.4.4. Then when ℓ is large

$$D_{r,1}(G_\ell, G_0) = \sum_{i=1}^{k_0} (\|a_i'\|_2^r (1/\ell)^r + |b_i'|(1/\ell)) \geq 1/\ell \sum_{i=1}^{k_0} |b_i'|$$

and thus

$$\frac{D_1(G_\ell, G_0)}{D_{r,1}(G_\ell, G_0)} \leq \frac{1/\ell}{1/\ell \sum_{i=1}^{k_0} |b_i'|} = \frac{1}{\sum_{i=1}^{k_0} |b_i'|} < \infty.$$

Moreover, as shown in the proof of Lemma 2.4.4,

$$\lim_{\ell \to \infty} \frac{V(p_{G_\ell}, p_{G_0})}{D_1(G_\ell, G_0)} = 0.$$

Combining the last two displays establishes

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{D_{r,1}(G, G_0)} = 0.$$

It can be verified that for large ℓ

$$W_r^r(G, G_0) \geq \frac{1}{8} \left(\min_{1 \leq i < j \leq k} \|\theta_i^0 - \theta_j^0\|_2 \right) \sum_{i=1}^{k} |p_i^\ell - p_i^0| = \frac{1}{8} \left(\min_{1 \leq i < j \leq k} \|\theta_i^0 - \theta_j^0\|_2 \right) \frac{1}{\ell} \sum_{i=1}^{k_0} |b_i'|.$$

The rest of the proof is similar to above to establish

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_G, P_{G_0})}{W_r^r(G, G_0)} = 0.$$

□

Proof of Lemma 2.4.9: It suffices to prove (2.12) since (2.11) is a direct consequence of (2.12).

Without loss of generality, assume $\theta_1^0 < \theta_2^0 < \ldots < \theta_{k_0}^0$. Let $\mathcal{N} = \bigcup_{i=1}^{k_0}(\theta_i^0 - \rho, \theta_i^0 + \rho)$, where $\rho = \frac{1}{4}\min_{1\le i<j\le k_0}|\theta_i^0 - \theta_j^0|$. Notice that for $x \in \mathbb{R}\backslash\mathcal{N}$, $f(x|\theta)$ as a function of θ is continuously differentiable on $(\theta_i^0 - \rho, \theta_i^0 + \rho)$ for each $i \in [k_0]$.

Suppose (2.12) is not true. Proceed exactly the same as the proof of Lemma 2.4.2 b) except the last paragraph to obtain a nonzero solution $(p_i^0 a_i, b_i : i \in [k_0])$ of (2.9a), (2.9b) with k, θ_i replaced by k_0, θ_i^0. For the uniform distribution family, one may argue that the nonzero solution has to satisfy

$$-p_i^0 a_i/\theta_i^0 + b_i = 0 \quad \forall i \in [k_0]. \tag{2.62}$$

Indeed, start from the rightmost interval that intersects with the support from only one mixture component, for $\mu - a.e.\ x \in (\theta_{k_0-1}^0, \theta_{k_0}^0)\backslash\mathcal{N} = [\theta_{k_0-1}^0 + \rho, \theta_{k_0}^0 - \rho]$

$$\begin{aligned}
0 &= \sum_{i=1}^{k_0}\left(p_i^0 a_i \frac{\partial}{\partial\theta} f(x|\theta_i^0) + b_i f(x|\theta_i^0)\right) \\
&= \sum_{i=1}^{k_0}\left(-p_i^0 a_i/\theta_i^0 + b_i\right) f(x|\theta_i^0) \\
&= \left(-p_{k_0}^0 a_{k_0}/\theta_{k_0}^0 + b_{k_0}\right)/\theta_{k_0}^0,
\end{aligned}$$

which implies $-p_{k_0}^0 a_{k_0}/\theta_{k_0}^0 + b_{k_0} = 0$. Repeat the above argument on interval $(\theta_{k_0-2}^0, \theta_{k_0-1}^0)$, ..., (θ_1^0, θ_2^0), $(0, \theta_1^0)$ and (2.62) is established.

Combining (2.62) with the fact that some of the a_i or b_i is non-zero, it follows that $|a_\alpha| > 0$ for some $\alpha \in [k_0]$. When ℓ is sufficiently large, $\theta_i^\ell, \eta_i^\ell \in (\theta_i^0 - \rho, \theta_i^0 + \rho)$. For sufficiently large ℓ

$$\begin{aligned}
&\frac{2V(P_{G_\ell}, P_{H_\ell})}{D_1(G_\ell, H_\ell)} \\
\ge &\frac{1}{D_1(G_\ell, H_\ell)}\int_{\min\{\theta_\alpha^\ell, \eta_\alpha^\ell\}}^{\max\{\theta_\alpha^\ell, \eta_\alpha^\ell\}} |p_{G_\ell}(x) - p_{H_\ell}(x)|\, dx \\
\overset{(*)}{=} &\frac{1}{D_1(G_\ell, H_\ell)}\int_{\min\{\theta_\alpha^\ell, \eta_\alpha^\ell\}}^{\max\{\theta_\alpha^\ell, \eta_\alpha^\ell\}} \left|\frac{(\pi_\alpha^\ell \mathbf{1}(\theta_\alpha^\ell < \eta_\alpha^\ell) + p_\alpha^\ell \mathbf{1}(\theta_\alpha^\ell \ge \eta_\alpha^\ell))}{\max\{\theta_\alpha^\ell, \eta_\alpha^\ell\}} + \sum_{i=\alpha+1}^{k_0}\frac{p_i^\ell}{\theta_i^\ell} - \sum_{i=\alpha+1}^{k_0}\frac{p_i^0}{\theta_i^0}\right| dx \\
\overset{(**)}{=} &\frac{1}{D_1(G_\ell, H_\ell)}|\theta_\alpha^\ell - \eta_\alpha^\ell|\left|(\pi_\alpha^\ell \mathbf{1}(\theta_\alpha^\ell < \eta_\alpha^\ell) + p_\alpha^\ell \mathbf{1}(\theta_\alpha^\ell \ge \eta_\alpha^\ell))/\max\{\theta_\alpha^\ell, \eta_\alpha^\ell\} + \sum_{i=\alpha+1}^{k_0}\frac{p_i^\ell}{\theta_i^\ell} - \sum_{i=\alpha+1}^{k_0}\frac{p_i^0}{\theta_i^0}\right| \\
\to &|a_\alpha|\frac{p_\alpha^0}{\theta_\alpha^0} > 0,
\end{aligned}$$

where the step $(*)$ follows from carefully examining the support of $f(x|\theta)$, the step $(**)$ follows from the integrand is a constant, and the last step follows from (2.60). The last display contradicts

with the choice of G_ℓ, H_ℓ, which satisfies $\frac{V(P_{G_\ell}, P_{H_\ell})}{D_1(G_\ell, H_\ell)} \to 0$.

□

Proof of Lemma 2.4.11: Without loss of generality, assume $\xi_1^0 \leq \xi_2^0 \leq \ldots \leq \xi_{k_0}^0$. Let $\mathcal{N} = \bigcup_{i=1}^{k_0} \{\xi_i^0\}$. Notice that for $x \in \mathbb{R} \backslash \mathcal{N}$, $f(x|\theta)$ as a function of θ is differentiable at $\theta_i^0 = (\xi_i^0, \sigma_i^0)$ for each $i \in [k_0]$.

Suppose (2.11) is not true. Proceed exactly the same as the proof of Lemma 2.4.2 a) except the last paragraph to obtain a nonzero solution $(p_i^0 a_i, b_i : i \in [k_0])$ of (2.9a), (2.9b) with k, θ_i replaced by k_0, θ_i^0. Write the two-dimensional vector a_i as $a_i = (a_i^{(\xi)}, a_i^{(\sigma)})$. For the location-scale exponential distribution, one may argue that the nonzero solution has to satisfy

$$a_i^{(\sigma)} = 0, \quad p_i^0 a_i^{(\xi)} / \sigma_i^0 + b_i = 0, \quad \forall i \in [k_0]. \tag{2.63}$$

Indeed, let $\bigcup_{i=1}^{k_0} \{\xi_i^0\} = \{\xi_1', \xi_2', \ldots, \xi_{k'}'\}$ with $\xi_1' < \xi_2' < \ldots < \xi_{k'}'$ where k' is the number of distinct elements. Define $I'(\xi) = \{i \in [k_0] : \xi_i^0 = \xi\}$. Then for $\mu - a.e.\ x \in \mathbb{R} \backslash \mathcal{N}$

$$\begin{aligned}
0 =& \sum_{i=1}^{k_0} \left(p_i^0 \langle a_i, \nabla_{(\xi,\sigma)} f(x|\xi_i^0, \sigma_i^0) \rangle + b_i f(x|\xi_i^0, \sigma_i^0) \right) \\
=& \sum_{j=1}^{k'} \sum_{i \in I'(\xi_j')} \left(p_i^0 \langle a_i, \nabla_{(\xi,\sigma)} f(x|\xi_j', \sigma_i^0) \rangle + b_i f(x|\xi_j', \sigma_i^0) \right) \\
=& \sum_{j=1}^{k'} \sum_{i \in I'(\xi_j')} \left(p_i^0 a_i^{(\xi)} \frac{1}{\sigma_i^0} + p_i^0 a_i^{(\sigma)} \frac{x - \xi_i^0 - \sigma_i^0}{(\sigma_i^0)^2} + b_i \right) f(x|\xi_j', \sigma_i^0).
\end{aligned}$$

Start from the leftmost interval that intersects with the support from only one mixture component, for $\mu - a.e.\ x \in (\xi_1', \xi_2') \backslash \mathcal{N} = [\xi_1' + \rho, \xi_2' - \rho]$,

$$\begin{aligned}
0 =& \sum_{i \in I'(\xi_1')} \left(p_i^0 a_i^{(\xi)} \frac{1}{\sigma_i^0} + p_i^0 a_i^{(\sigma)} \frac{x - \xi_i^0 - \sigma_i^0}{(\sigma_i^0)^2} + b_i \right) f(x|\xi_1', \sigma_i^0) \\
=& \sum_{i \in I'(\xi_1')} \left(p_i^0 a_i^{(\xi)} \frac{1}{\sigma_i^0} + p_i^0 a_i^{(\sigma)} \frac{x - \xi_i^0 - \sigma_i^0}{(\sigma_i^0)^2} + b_i \right) \exp\left(\frac{\xi_1'}{\sigma_i^0} \right) \exp\left(-\frac{x}{\sigma_i^0} \right).
\end{aligned}$$

Since σ_i^0 for $i \in I'(\xi_1')$ are all distinct, by Lemma 2.10.4 a)

$$a_i^{(\sigma)} = 0, \quad p_i^0 a_i^{(\xi)} / \sigma_i^0 + b_i = 0, \quad \forall i \in I'(\xi_1').$$

Repeat the above argument on interval $(\xi_2', \xi_3'), \ldots, (\xi_{k'-1}', \xi_{k'}'), (\xi_{k'}', \infty)$ and (2.63) is established.

Since at least one of a_i or b_i is not zero, from (2.63) it is clear that at least one of $\{b_i\}_{i=1}^{k_0}$

is not zero. Then by $\sum_{i=1}^{k_0} b_i = 0$ at least one of b_i is positive. By (2.63) at least one of $a_i^{(\xi)}$ is negative. Let $\alpha \in \underset{i\in\{j\in[k_0]:a_j^{(\xi)}<0\}}{\arg\max} a_i^{(\xi)}$. That is $a_\alpha^{(\xi)}$ is a largest negative one among $\{a_i^{(\xi)}\}_{i\in[k_0]}$. Let $\rho = \frac{1}{2}\min_{1\le i<j\le k'} |\xi_i' - \xi_j'|$ to be half of the smallest distance among different $\{\xi_i'\}_{i=1}^{k'}$. By subsequence argument if necessary, we require for any $i \in [k_0]$, $\xi_i^\ell \in (\xi_i^0 - \rho, \xi_i^0 + \rho)$.

Let $I(\alpha) = \{i \in [k_0] | \xi_i^0 = \xi_\alpha^0\}$ to be the set of indices for those sharing the same ξ_i^0 as ξ_α^0. We now consider subsequences such that ξ_i^ℓ for $i \in I(\alpha)$ satisfies finer properties as follows. Divide the index set $I(\alpha)$ into three subsets, $J(\alpha) := \{i \in I(\alpha) | a_i^{(\xi)} = a_\alpha^{(\xi)}\}$, $J_<(\alpha) := \{i \in I(\alpha) | a_i^{(\xi)} < a_\alpha^{(\xi)}\}$ and $J_>(\alpha) := \{i \in I(\alpha) | a_i^{(\xi)} > a_\alpha^{(\xi)}\}$. Note $J(\alpha)$ is the index set for those sharing the same ξ_i^0 as ξ_α^0 and sharing the same $a_i^{(\xi)}$ as $a_\alpha^{(\xi)}$ (so their $a_i^{(\xi)}$ are also largest negative ones among $\{a_i^{(\xi)}\}_{i\in[k_0]}$), while $J_>(\alpha)$ corresponds for indices i for which $\xi_i^0 = \xi_\alpha^0$ and $a_i^{(\xi)} \ge 0$, and $J_<(\alpha)$ corresponds for indices i for which $\xi_i^0 = \xi_\alpha^0$ and $a_i^{(\xi)} < a_\alpha^{(\xi)}$. To be clear, the two subsets $J_{<\alpha}$ and $J_{>\alpha}$ may be empty, but J_α is non-empty by our definition.

For any $i \in J_<(\alpha)$, $j \in J(\alpha)$

$$\frac{\xi_i^\ell - \xi_\alpha^0}{D_\ell(G_\ell, G_0)} \to a_i^{(\xi)} < a_\alpha^{(\xi)} \leftarrow \frac{\xi_j^\ell - \xi_\alpha^0}{D_\ell(G_\ell, G_0)}.$$

Then for large ℓ, $\xi_i^\ell < \xi_j^\ell$ for any $i \in J_<(\alpha)$ and $j \in J(\alpha)$. Similarly for large ℓ, $\xi_j^\ell < \xi_k^\ell$ for any $j \in J(\alpha)$ and $k \in J_>(\alpha)$. Thus by subsequence argument if necessary, we require ξ_i^ℓ additionally satisfy the conditions specified in the last two sentences for all ℓ.

Consider $\max_{j\in J(\alpha)}\{\xi_j^\ell\}$ and there exists $\bar\alpha \in J(\alpha)$ such that $\xi_{\bar\alpha}^\ell = \max_{j\in J(\alpha)}\{\xi_j^\ell\}$ for infinitely many ℓ since $J(\alpha)$ has finite cardinality. By subsequence argument if necessary, we require $\xi_{\bar\alpha}^\ell = \max_{j\in J(\alpha)}\{\xi_j^\ell\}$ for all ℓ. Moreover, since $a_{\bar\alpha}^\xi = a_\alpha^\xi < 0$ we may further require $\xi_{\bar\alpha}^\ell < \xi_\alpha^0$ for all ℓ. Finally, for each $k \in J_>(\alpha)$ such that $a_k^{(\xi)} > 0$, we may further require $\xi_k^\ell > \xi_\alpha^0$ for all ℓ by subsequences.

To sum up, $\{\xi_i^\ell\}$ for $i \in I(\alpha)$ satisfy:

$$\begin{cases} \xi_i^\ell \le \xi_{\bar\alpha}^\ell < \xi_\alpha^0, & \forall \ell, \quad \forall\, i \in J_<(\alpha) \bigcup J(\alpha) \\ \xi_i^\ell > \xi_{\bar\alpha}^\ell, & \forall \ell, \quad \forall\, i \in J_>(\alpha) \\ \xi_i^\ell > \xi_\alpha^0, & \forall \ell, \forall i \in J_>(\alpha) \text{ and } a_i^{(\xi)} > 0 \end{cases} . \tag{2.64}$$

Let $\bar\xi^\ell = \min\left\{ \min_{i\in\{j\in I(\alpha):a_j^{(\xi)}=0\}} \xi_i^\ell, \xi_\alpha^0 \right\}$ with the convention that the minimum over an empty set is ∞. Then $\bar\xi^\ell \le \xi_\alpha^0$ and $\bar\xi^\ell \to \xi_\alpha^0$. Moreover, by property (2.64), $\bar\xi^\ell > \xi_{\bar\alpha}^\ell$. Thus on $(\xi_{\bar\alpha}^\ell, \bar\xi^\ell)$, 1) for any $i > \max I(\alpha)$, $f(x|\xi_i^\ell, \sigma_i^\ell) = 0 = f(x|\xi_i^0, \sigma_i^0)$ since $\xi_i^\ell, \xi_i^0 \ge \xi_\alpha^0 \ge \bar\xi^\ell$; 2) for $i \in J_>(\alpha)$, $f(x|\xi_i^\ell, \sigma_i^\ell) = 0$ due to $\xi_i^\ell \ge \bar\xi^\ell$ due to (2.64); 3) for $i \in I(\alpha)$, $f(x|\xi_i^0, \sigma_i^0) = 0$ since $\xi_i^0 = \xi_\alpha^0 \ge \bar\xi^\ell$.

Then

$$
\begin{aligned}
&\frac{2V(P_{G_\ell}, P_{G_0})}{D_1(G_\ell, G_0)} \\
\geq &\frac{1}{D_1(G_\ell, G_0)} \int_{\xi^\ell_{\bar\alpha}}^{\bar\xi^\ell} |p_{G_\ell}(x) - p_{G_0}(x)|\, dx \\
= &\frac{1}{D_1(G_\ell, G_0)} \int_{\xi^\ell_{\bar\alpha}}^{\bar\xi^\ell} \Bigg| \sum_{i \in J_<(\alpha) \bigcup J(\alpha)} p_i^\ell \frac{1}{\sigma_i^\ell} \exp\left(-\frac{x-\xi_\alpha^\ell}{\sigma_i^\ell}\right) \\
&\quad + \sum_{i < \min I(\alpha)} \left(p_i^\ell \frac{1}{\sigma_i^\ell} \exp\left(-\frac{x-\xi_i^\ell}{\sigma_i^\ell}\right) - p_i^0 \frac{1}{\sigma_i^0} \exp\left(-\frac{x-\xi_i^0}{\sigma_i^0}\right)\right) \Bigg| dx. \qquad (2.65)
\end{aligned}
$$

Denote the integrand (including the absolute value) in the preceding display by $A_\ell(x)$. Then as a function on $[\xi_\alpha^0 - \rho, \xi_\alpha^0]$, $A_\ell(x)$ converges unifomrly to $\sum_{i \in J_<(\alpha) \bigcup J(\alpha)} p_i^0 \frac{1}{\sigma_i^0} \exp\left(-\frac{x-\xi_\alpha^0}{\sigma_i^0}\right) := B(x)$. Since $B(x)$ is positive and continuous on compact interval $[\xi_\alpha^0 - \rho, \xi_\alpha^0]$, for large ℓ

$$
|A_\ell(x) - B(x)| \leq \frac{1}{\ell} \leq \frac{1}{2} \min B(x) \leq \frac{1}{2} B(x), \quad \forall x \in [\xi_\alpha^0 - \rho, \xi_\alpha^0],
$$

which yields

$$
A_\ell(x) \geq \frac{1}{2} B(x) \geq \frac{1}{2} p_{\bar\alpha}^0 \frac{1}{\sigma_{\bar\alpha}^0} \exp\left(-\frac{x-\xi_\alpha^0}{\sigma_{\bar\alpha}^0}\right) \geq \frac{1}{2} p_{\bar\alpha}^0 \frac{1}{\sigma_{\bar\alpha}^0}, \quad \forall x \in [\xi_\alpha^0 - \rho, \xi_\alpha^0].
$$

Plug the preceding display into (2.65), one obtains for large ℓ,

$$
\begin{aligned}
\frac{2V(P_{G_\ell}, P_{H_\ell})}{D_1(G_\ell, H_\ell)} \geq &\frac{1}{D_1(G_\ell, G_0)} \int_{\xi^\ell_{\bar\alpha}}^{\bar\xi^\ell} \frac{1}{2} p_{\bar\alpha}^0 \frac{1}{\sigma_{\bar\alpha}^0} dx \\
= &\left(\frac{\xi_\alpha^0 - \xi_{\bar\alpha}^\ell}{D_1(G_\ell, G_0)} - \frac{\xi_\alpha^0 - \bar\xi^\ell}{D_1(G_\ell, G_0)} \right) \frac{1}{2} p_{\bar\alpha}^0 \frac{1}{\sigma_{\bar\alpha}^0} \\
\to &(-a_{\bar\alpha}^{(\xi)} - 0) \frac{1}{2} p_{\bar\alpha}^0 \frac{1}{\sigma_{\bar\alpha}^0} > 0 \qquad (2.66)
\end{aligned}
$$

where the convergence in the last step is due to (2.14). (2.66) contradicts with the choice of G_ℓ, which satisfies $\frac{V(P_{G_\ell}, P_{G_0})}{D_1(G_\ell, G_0)} \to 0$.

□

Proof of Lemma 2.4.12: Take $\tilde f(x) = \max_{i \in [k_0]} \bar f(x) \sqrt{f(x|\theta_i^0)}$. Then $\tilde f(x)$ is μ-integrable by

Cauchy-Schwarz inequality. Moreover for any $i \in [k_0]$ and any $0 < \Delta \leq \gamma_0$

$$\left| \frac{f(x|\theta_i^0 + a_i\Delta) - f(x|\theta_i^0)}{\Delta} \right| \leq \tilde{f}(x) \quad \mu - a.e.\ x \in \mathfrak{X}.$$

Then by Lemma 2.4.14 b) $(a_1, b_1, \ldots, a_{k_0}, b_{k_0})$ is a nonzero solution of the system of equations (2.9a), (2.9b).

Let $a_i' = \frac{a_i/p_i^0}{\sum_{i=1}^{k_0}\left(\|a_i/p_i^0\|_2 + |b_i|\right)}$ and $b_i' = \frac{b_i}{\sum_{i=1}^{k_0}\left(\|a_i/p_i^0\|_2 + |b_i|\right)}$. Then a_i' and b_i' satisfy

$$\sum_{i=1}^{k_0} \left(\|a_i'\|_2 + |b_i'|\right) = 1$$

and $(p_1^0 a_1',\ b_1', \ldots, p_{k_0}^0 a_{k_0}',\ b_{k_0}')$ is also a nonzero solution of (2.9a), (2.9b) with k, θ_i replaced respectively by k_0, θ_i^0. Let $G_\ell = p_i^\ell \delta_{\theta_i^\ell}$ with $p_i^\ell = p_i^0 + b_i' \frac{1}{\ell}$ and $\theta_i^\ell = \theta_i^0 + \frac{1}{\ell} a_i'$ for $1 \leq i \leq k_0$. When ℓ is large, $0 < p_i^\ell < 1$ and $\theta_i^\ell \in \Theta$ since $0 < p_i^0 < 1$ and $\theta_i^0 \in \Theta^\circ$. Moreover, $\sum_{i=1}^{k_0} p_i^\ell = 1$ since $\sum_{i=1}^{k_0} b_i' = 0$. Then $G_\ell \in \mathcal{E}_{k_0}(\Theta)$ and $G_\ell \neq G_0$ since at least one of a_i' or b_i' is nonzero. When ℓ is large $D_1(G_\ell, G_0) = \sum_{i=1}^{k_0} \left(\|\theta_i^\ell - \theta_i^0\|_2 + |p_i^\ell - p_i^0|\right) = \frac{1}{\ell}$. Thus when ℓ is large

$$\begin{aligned}
&\frac{2h^2(P_{G_\ell}, P_{G_0})}{D_1^2(G_\ell, G_0)} \\
&= \int_S \left| \frac{p_{G_\ell}(x) - p_{G_0}(x)}{D_1(G_\ell, G_0)} \frac{1}{\sqrt{p_{G_\ell}(x)} + \sqrt{p_{G_0}(x)}} \right|^2 \mu(dx) \\
&= \int_{S\setminus\mathcal{N}} \left| \left(\sum_{i=1}^{k_0} p_i^\ell \frac{f(x|\theta_i^\ell) - f(x|\theta_i^0)}{1/\ell} + \sum_{i=1}^{k_0} b_i' f(x|\theta_i^0) \right) \frac{1}{\sqrt{p_{G_\ell}(x)} + \sqrt{p_{G_0}(x)}} \right|^2 \mu(dx).
\end{aligned}$$

The integrand of the last integral is bounded by

$$\begin{aligned}
&\left| \sum_{i=1}^{k_0} \frac{p_i^\ell}{\sqrt{p_i^0}} \frac{f(x|\theta_i^\ell) - f(x|\theta_i^0)}{1/\ell \times \sqrt{f(x|\theta_i^0)}} + \sum_{i=1}^{k_0} \frac{b_i'}{\sqrt{p_i^0}} \sqrt{f(x|\theta_i^0)} \right|^2 \\
\leq & 2k_0 \sum_{i=1}^{k_0} \frac{(p_i^\ell)^2}{p_i^0} \left| \frac{f(x|\theta_i^\ell) - f(x|\theta_i^0)}{1/\ell \times \sqrt{f(x|\theta_i^0)}} \right|^2 + 2k_0 \sum_{i=1}^{k_0} \frac{(b_i')^2}{p_i^0} f(x|\theta_i^0) \\
\leq & 2k_0 \sum_{i=1}^{k_0} \frac{1}{p_i^0} \left(\frac{1/p_i^0}{\sum_{i=1}^{k_0} \left(\|a_i/p_i^0\|_2 + |b_i|\right)} \right)^2 \bar{f}^2(x) + 2k_0 \sum_{i=1}^{k_0} \frac{(b_i')^2}{p_i^0} f(x|\theta_i^0),
\end{aligned}$$

which is integrable w.r.t. to μ on $S\backslash\mathcal{N}$. Here the last inequalities follows from

$$\left|\frac{f(x|\theta_i^\ell)-f(x|\theta_i^0)}{1/\ell\times\sqrt{f(x|\theta_i^0)}}\right| = \left|\frac{f(x|\theta_i^0+\frac{\|a_i'\|_2}{\|a_i\|_2}a_i\Delta)-f(x|\theta_i^0)}{\Delta\sqrt{f(x|\theta_i^0)}}\right| \leq \frac{\|a_i'\|_2}{\|a_i\|_2}\bar{f}(x).$$

Then by the dominated convergence theorem

$$\begin{aligned}\lim_{\ell\to\infty}\frac{2h^2(P_{G_\ell}(x),P_{G_0}(x))}{D_1^2(G_\ell,G_0)} &= \int_{S\backslash\mathcal{N}}\left|\left(\sum_{i=1}^{k_0}p_i^0\langle a_i',\nabla_\theta f(x|\theta_i^0)\rangle+\sum_{i=1}^{k_0}b_i'f(x|\theta_i^0)\right)\frac{1}{2\sqrt{p_{G_0}(x)}}\right|^2\mu(dx)\\ &=0.\end{aligned}$$

The proof is completed by

$$\lim_{r\to 0}\inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}}\frac{h(P_G,P_H)}{D_1(G,H)} \leq \liminf_{\substack{G\overset{W_1}{\to}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}}\frac{h(P_G,P_{G_0})}{D_1(G,G_0)}.$$

□

2.10.2 Proofs in Section 2.4.2

Proof of Lemma 2.4.14:

a) For any $x\in\mathfrak{X}\backslash\mathcal{N}$, $\nabla_\theta\tilde{f}(x|\theta_i)=g(\theta_i)\nabla_\theta f(x|\theta_i)+f(x|\theta_i)\nabla_\theta g(x|\theta_i)$. Then $(\tilde{a}_1,\tilde{b}_1,\ldots,\tilde{a}_k,\tilde{b}_k)$ is a solution of (2.9a) with f replaced by $\tilde{f}$ if and only if $(a_1,b_1,\ldots,a_k,b_k)$ with $a_i=g(\theta_i)\tilde{a}_i$ and $b_i=\langle\tilde{a}_i,\nabla_\theta g(\theta_i)\rangle+\tilde{b}_i g(\theta_i)$ is a solution of (2.9a). We can write $\tilde{a}_i=a_i/g(\theta_i)$ and $\tilde{b}_i=(b_i-\langle a_i,\nabla_\theta g(\theta_i)\rangle/g(\theta_i))/g(\theta_i)$. Thus $(\tilde{a}_1,\tilde{b}_1,\ldots,\tilde{a}_k,\tilde{b}_k)$ is zero if and only if $(a_1,b_1,\ldots,a_k,b_k)$ is zero.

b) Under the conditions, by Dominated Convergence Theorem

$$\int_{\mathfrak{X}\backslash\mathcal{N}}\langle a_i,\nabla_\theta f(x|\theta_i)\rangle d\mu=\langle a_i,\nabla_\theta\int_{\mathfrak{X}\backslash\mathcal{N}}f(x|\theta)d\mu\rangle\Big|_{\theta=\theta_i}=0.$$

where the last step follows from $\mu(\mathcal{N})=0$ and the fact that $f(x|\theta)$ is a density w.r.t. μ. Thus for $(a_1,b_1,\ldots,a_k,b_k)$ any solution of (2.9a),

$$\sum_{i=1}^k b_i=\int_{\mathfrak{X}\backslash\mathcal{N}}\sum_{i=1}^k\left(\langle a_i,\nabla_\theta f(x|\theta_i)\rangle+b_i f(x|\theta_i)\right)d\mu=0.$$

So $(a_1,b_1,\ldots,a_k,b_k)$ is also a solution of the system (2.9a),(2.9b).

It remains to show (2.20) is equivalent to the same conditions on $\tilde{f}$. Suppose (2.20) is true. Then there exists small enough $\tilde{\gamma}(\theta_i, a_i) < \gamma(\theta_i, a_i)$ such that for $0 < \Delta \leq \tilde{\gamma}(\theta_i, a_i)$

$$\begin{aligned}
&\left|\frac{\tilde{f}(x|\theta_i + a_i\Delta) - \tilde{f}(x|\theta_i)}{\Delta}\right| \\
\leq& g(\theta_i + a_i\Delta)\left|\frac{f(x|\theta_i + a_i\Delta) - f(x|\theta_i)}{\Delta}\right| + \left|\frac{g(\theta_i + a_i\Delta) - g(\theta_i)}{\Delta}\right| f(x|\theta_i) \\
\leq& C(g, \theta_i, a_i, \tilde{\gamma}(\theta_i, a_i))(\bar{f}(x|\theta_i, a_i) + f(x|\theta_i)) \quad \mu - a.e.\ \mathfrak{X}
\end{aligned}$$

and thus one can take μ-integrable $\bar{f}_1(x|\theta_i, a_i) = C(g, \theta_i, a_i, \tilde{\gamma}(\theta_i, a_i))(\bar{f}(x|\theta_i, a_i) + f(x|\theta_i))$. The reverse direction follows similarly.

c) It's a direct consequence from parts a) and b).

□

Proof of Lemma 2.4.16: Notice that $f(x|\theta)$ is continuously differentiable at every $\theta \in \Theta^\circ$ when fixing any $x \in \mathfrak{X}$. By Lemma 2.10.2 and Lemma 2.4.14 c), (2.9a) has the same solutions as the system (2.9a),(2.9b).

It's obvious that a) implies b) and that c) implies d). That a) implies c) and that b) implies d) follow from $V(p_G, p_{G_0}) \leq h(p_G, p_{G_0})$. e) implies a) follows from Lemma 2.4.2 b). It remains to prove d) implies e).

Suppose d) holds and the system of equations (2.9a), (2.9b) with k, θ_i replaced respectively by k_0, θ_i^0 has a nonzero solution $(a_1, b_1, \ldots, a_{k_0}, b_{k_0})$. By Lemma 2.10.2, the condition d) of Lemma 2.4.12 is satisfied with $\gamma_0 = \min_{i\in[k_0]} \gamma(\theta_i^0, a_i)$ and $\bar{f}(x) = \max_{i\in[k_0]} \bar{f}(x|\theta_i^0, a_i)$. Thus by Lemma 2.4.12, d) does not hold. This is a contradiction and thus d) implies e). □

Lemma 2.10.2. *Let $f(x|\theta)$ be the density of a full rank exponential family in canonical form specified as in Lemma 2.4.16. Then for any $\theta \in \Theta^\circ$ and $a \in \mathbb{R}^q$ there exists $\gamma(\theta, a) > 0$ such that for any $0 < \Delta \leq \gamma(\theta, a)$,*

$$\left|\frac{f(x|\theta + a\Delta) - f(x|\theta)}{\Delta\sqrt{f(x|\theta)}}\right| \leq \bar{f}(x|\theta, a) \quad \forall x \in S = \{x|f(x|\theta) > 0\}$$

with $\int_{\mathfrak{X}} \bar{f}^2(x|\theta, a)d\mu < \infty$ and

$$\left|\frac{f(x|\theta + a\Delta) - f(x|\theta)}{\Delta}\right| \leq \tilde{f}(x|\theta, a) \quad \forall x \in \mathfrak{X}$$

with $\int_{\mathfrak{X}} \tilde{f}(x|\theta, a)d\mu < \infty$. Here $\gamma(\theta, a)$, $\bar{f}(x|\theta, a)$ and $\tilde{f}(x|\theta, a)$ depend on θ and a.

Proof of Lemma 2.10.2: Let $\gamma > 0$ be such that the line segment between $\theta - a\gamma$ and $\theta + a\gamma$ lie in Θ and $\int_{\mathfrak{X}} e^{4\gamma^T T(x)} f(x|\theta) d\mu < \infty$, $\int_{\mathfrak{X}} e^{-4\gamma^T T(x)} f(x|\theta) d\mu < \infty$ due to the fact that the moment generating function exists in a neighborhood of origin for any given $\theta \in \Theta^\circ$. Then for $\Delta \in (0, \gamma]$ and for any $x \in S$

$$
\begin{aligned}
&\left| \frac{f(x|\theta + a\Delta) - f(x|\theta)}{\Delta \sqrt{f(x|\theta)}} \right| \\
=&\sqrt{f(x|\theta)} \left| \frac{\exp(\langle a\Delta, T(x)\rangle - (A(\theta + a\Delta) - A(\theta))) - 1}{\Delta} \right| \\
\overset{(*)}{\leq}&\sqrt{f(x|\theta)} \left| \langle a, T(x)\rangle - \frac{A(\theta + a\Delta) - A(\theta)}{\Delta} \right| e^{\langle a\Delta, T(x)\rangle - (A(\theta + a\Delta) - A(\theta))} \\
\leq&\sqrt{f(x|\theta)} \left(|\langle a, T(x)\rangle| + \|a\|_2 \max_{\Delta \in [0,\gamma]} \|\nabla_\theta A(\theta + a\Delta)\|_2 \right) e^{\Delta |\langle a, T(x)\rangle|} \max_{\Delta \in [0,\gamma]} e^{-(A(\theta + a\Delta) - A(\theta))} \\
\leq&\sqrt{f(x|\theta)} \frac{1}{\gamma} e^{\gamma |\langle a, T(x)\rangle| + \gamma \|a\|_2 \max_{\Delta \in [0,\gamma]} \|\nabla_\theta A(\theta + a\Delta)\|_2} \; e^{\gamma |\langle a, T(x)\rangle|} \max_{\Delta \in [0,\gamma]} e^{-(A(\theta + a\Delta) - A(\theta))} \\
=&C(\gamma, a, \theta) \sqrt{f(x|\theta)} e^{2\gamma |\langle a, T(x)\rangle|} \\
\leq&\sqrt{C^2(\gamma, a, \theta) f(x|\theta) \left(e^{4\gamma \langle a, T(x)\rangle} + e^{-4\gamma \langle a, T(x)\rangle}\right)}, \qquad (2.67)
\end{aligned}
$$

where step $(*)$ follows from $|e^t - 1| \leq |t| e^t$. Then the the first conclusion holds with

$$
\bar{f} = \sqrt{C^2(\gamma, a, \theta) f(x|\theta) \left(e^{4\gamma \langle a, T(x)\rangle} + e^{-4\gamma \langle a, T(x)\rangle}\right)}.
$$

Take $\tilde{f}(x) = \bar{f}(x) \sqrt{f(x|\theta)}$ and by Cauchy–Schwarz inequality $\int_{\mathfrak{X}} \tilde{f}(x) d\mu \leq \int_{\mathfrak{X}} \bar{f}^2(x) d\mu < \infty$. Moreover by (2.67)

$$
\left| \frac{f(x|\theta_i^0 + \Delta a_i) - f(x|\theta_i^0)}{\Delta} \right| \leq \tilde{f}(x) \quad \forall x \in \mathfrak{X}.
$$

□

Proof of Lemma 2.4.17: Consider $\tilde{f}(x|\eta) := f(x|\theta)$ to be the same kernel but under the new parameter $\eta = \eta(\theta)$. Note $\{\tilde{f}(x|\eta)\}_{\eta \in \Theta}$ with $\Xi := \eta(\Theta)$ is the canonical parametrization of the same exponential family. Write $\eta_i^0 = \eta(\theta_i^0)$. Since $J_\eta(\theta) = (\frac{\partial \eta^{(i)}}{\partial \theta^{(j)}}(\theta))_{ij}$ exists at θ_i^0 and at those points,

$$
\nabla_\theta f(x|\theta_i^0) = \left(J_\eta(\theta_i^0)\right)^T \nabla_\eta \tilde{f}(x|\eta_i^0), \quad \forall i \in [k_0]
$$

and thus

$$
\sum_{i=1}^{k_0} \left(\langle a_i, \nabla_\theta f(x|\theta_i^0)\rangle + b_i f(x|\theta_i^0) \right) = \sum_{i=1}^{k_0} \left(\langle J_\eta(\theta_i^0) a_i, \nabla_\eta \tilde{f}(x|\eta_i^0)\rangle + b_i \tilde{f}(x|\eta_i^0) \right). \qquad (2.68)
$$

Then (2.9a), (2.9b) with k, θ_i replaced respectively by k_0, θ_i^0 has only the zero solution if and only if (2.9a), (2.9b) with k, θ_i, f replaced respectively by $k_0, \eta_i^0, \tilde{f}$ has only the zero solution.

Suppose that $(a_1, b_1, \ldots, a_{k_0}, b_{k_0})$ is a solution of (2.9a) with k, θ_i replaced respectively by k_0, θ_i^0. Then by (2.68) $(\tilde{a}_1, \tilde{b}_1, \ldots, \tilde{a}_{k_0}, \tilde{b}_{k_0})$ with $\tilde{a}_i = J_\eta(\theta_i^0)a_i$, $\tilde{b}_i = b_i$ is a solution of (2.9a) with k, θ_i, f replaced respectively by $k_0, \eta_i^0, \tilde{f}$. Then by Lemma 2.4.16, it necessarily has $\sum_{i=1}^{k_0} b_i = \sum_{i=1}^{k_0} \tilde{b}_i = 0$. That is, $(a_1, b_1, \ldots, a_{k_0}, b_{k_0})$ is a solution of the system of equations (2.9a), (2.9b) with k, θ_i replaced respectively by k_0, θ_i^0. As a result, with k, θ_i replaced respectively by k_0, θ_i^0, (2.9a) has the same solutions as the system (2.9a),(2.9b).

The rest of the proof is completed by appealing to Lemma 2.5.6 and Lemma 2.4.16.

□

2.10.3 Auxiliary Lemmas

Lemma 2.10.3. *Consider $g(x)$ on $\mathbb{R}^d$ is a function with its gradient $\nabla g(x)$ continuous in a neighborhood of x_0.*

a) Then when $x \to x_0$ and $y \to x_0$

$$|g(x) - g(y) - \langle \nabla g(x_0), x - y \rangle| = o(\|x - y\|_2)$$

b) If in addition, the Hessian $\nabla^2 g(x)$ is continuous in a neighborhood of x_0. Then for any x, y in a closed ball B of x_0 contained in that neighborhood,

$$\begin{aligned}
&|g(x) - g(y) - \langle \nabla g(x_0), x - y \rangle| \\
\leq &\int_0^1 \int_0^1 \|\nabla^2 g(x_0 + s(y + t(x - y) - x_0))\|_2 ds dt \, \|x - y\|_2 \max\{\|x - x_0\|_2, \|y - x_0\|_2\} \\
\leq &d \sum_{1 \leq i,j \leq d} \int_0^1 \int_0^1 \left| \frac{\partial^2 g}{\partial x^{(i)} x^{(j)}}(x_0 + s(y + t(x - y) - x_0)) \right| ds dt \times \\
&\|x - y\|_2 \max\{\|x - x_0\|_2, \|y - x_0\|_2\}.
\end{aligned}$$

Moreover

$$|g(x) - g(y) - \langle \nabla g(x_0), x - y \rangle| \leq L \|x - y\|_2 \max\{\|x - x_0\|_2, \|y - x_0\|_2\}.$$

where $L = \sup_{x \in B} \|\nabla^2 g(x)\|_2 < \infty$.

Proof: a)

$$\begin{aligned}
&\lim_{x\neq y, x\to x_0, y\to x_0} \frac{|g(x)-g(y)-\langle \nabla g(x_0), x-y\rangle|}{\|x-y\|_2} \\
=&\lim_{x\neq y, x\to x_0, y\to x_0} \frac{|\langle \nabla g(\xi), x-y\rangle - \langle \nabla g(x_0), x-y\rangle|}{\|x-y\|_2} \\
\leq&\lim_{x\neq y, x\to x_0, y\to x_0} \|\nabla g(\xi) - \nabla g(x_0)\|_2 \\
=&0,
\end{aligned}$$

where the first step follows from mean value theorem with ξ lie in the line segment connecting x and y, the second step follows from Cauchy-Schwarz inequality, and the last step follows from the continuity of $\nabla g(x)$ and $\xi \to x_0$ when $x, y \to x_0$.

b) For $x \neq y$ in B specified in the statement,

$$\begin{aligned}
&\frac{|g(x)-g(y)-\langle \nabla g(x_0), x-y\rangle|}{\|x-y\|_2} \\
=&\frac{|\int_0^1 \langle \nabla g(y+t(x-y)), x-y\rangle dt - \langle \nabla g(x_0), x-y\rangle|}{\|x-y\|_2} \\
=&\frac{|\int_0^1\int_0^1 \langle\langle \nabla^2 g(x_0+s(y+t(x-y)-x_0)), y+t(x-y)-x_0\rangle, x-y\rangle ds dt|}{\|x-y\|_2} \\
\leq&\frac{\int_0^1\int_0^1 |\langle\langle \nabla^2 g(x_0+s(y+t(x-y)-x_0)), y+t(x-y)-x_0\rangle, x-y\rangle| ds dt}{\|x-y\|_2} \\
\leq&\int_0^1\int_0^1 \|\nabla^2 g(x_0+s(y+t(x-y)-x_0))\|_2 \|y+t(x-y)-x_0\|_2 ds dt \\
\leq&\int_0^1\int_0^1 \|\nabla^2 g(x_0+s(y+t(x-y)-x_0))\|_2 ds dt \ \max\{\|x-x_0\|_2, \|y-x_0\|_2\}, \quad (2.69)
\end{aligned}$$

where the first two equalities follows respectively form fundamental theorem of calculus for $\mathbb{R}$-valued functions and $\mathbb{R}^d$-valued functions. Moreover, observe for any matrix $A \in \mathbb{R}^{d\times d}$,

$$\|A\|_2 \leq \|A\|_F \leq d \max_{1\leq i,j\leq d} |A_{ij}| \leq d \sum_{1\leq i,j\leq d} |A_{ij}|$$

where $\|\cdot\|_F$ is the Frobenius norm. Apply the preceding display to (2.69),

$$\begin{aligned}
&\int_0^1\int_0^1 \|\nabla^2 g(x_0+s(y+t(x-y)-x_0))\|_2 ds dt \\
\leq& d \sum_{1\leq i,j\leq d} \int_0^1\int_0^1 \left| \frac{\partial^2 g}{\partial x^{(i)} x^{(j)}} (x_0+s(y+t(x-y)-x_0)) \right| ds dt
\end{aligned}$$

Following (2.69),

$$\frac{|g(x) - g(y) - \langle \nabla g(x_0), x - y \rangle|}{\|x - y\|_2} \leq L \max\{\|x - x_0\|_2, \|y - x_0\|_2\}.$$

□

Lemma 2.10.4. *Let k be a positive integer, $b_1 < \ldots < b_k$ be a sequence of real numbers and let μ be the Lebesgue measure on $\mathbb{R}$.*

a) Let $\{h_i(x)\}_{i=1}^k$ be a sequence of polynomials. Consider any nonempty interval I. Then

$$\sum_{i=1}^{k} h_i(x) e^{b_i x} = 0 \quad \mu - a.e.\ x \in I$$

implies $h_i(x) \equiv 0$ for any $i \in [k]$.

b) Let $\{h_i(x)\}_{i=1}^k$ be a sequence of functions, where each is of the form $\sum_{j=1}^{m_i} a_j x^{\gamma_j}$, i.e. a finite linear combination of power functions. Let $\{g_i(x)\}_{i=1}^k$ be another sequence of such functions. Consider any nonempty interval $I \subset (0, \infty)$. Then

$$\sum_{i=1}^{k} (h_i(x) + g_i(x) \ln(x)) e^{b_i x} = 0 \quad \mu - a.e.\ x \in I$$

implies when $x \neq 0$ $h_i(x) \equiv 0$ and $g_i(x) \equiv 0$ for any $i \in [k]$.

Proof:

a) Define $F(x) = \sum_{i=1}^k h_i(x) e^{b_i x}$. From the condition $F(x) = 0$ on a dense subset of I. Then $F(x) = 0$ on the closure of that subset, which contains I, since it is a continuous function on $\mathbb{R}$. Let $a \in I^\circ$ and consider its Taylor expansion $F(x) = \sum_{i=0}^{\infty} \frac{F^{(i)}(a)}{i!}(x - a)^i$ for any $x \in \mathbb{R}$. It follows from $F(x) = 0$ on I that $F^{(i)}(a) = 0$ for any $i \geq 0$. Thus $F(x) \equiv 0$ on $\mathbb{R}$. Then

$$0 = \lim_{x \to \infty} e^{-b_k x} F(x) = \lim_{x \to \infty} h_k(x).$$

This happen only when $h_k(x) \equiv 0$. Proceed in the same manner to show $h_i(x) \equiv 0$ for i from $k - 1$ to 1.

b) Define $H(x) = \sum_{i=1}^k (h_i(x) + g_i(x) \ln(x)) e^{b_i x}$. From the condition $H(x) = 0$ on a dense subset of I. Then $H(x) = 0$ on the closure of that subset excluding 0, which contains I, since it is a continuous function on $(0, \infty)$. Let $a_1 \in I^\circ$ and consider its Taylor expansion at a_1 $H(x) = \sum_{i=0}^{\infty} \frac{H^{(i)}(a_1)}{i!}(x - a_1)^i$ for $x \in (0, 2a_1)$, since the Taylor series of $\ln(x)$, x^γ at

a_1 converges respectively to $\ln(x)$, x^γ on $(0, 2a_1)$ for any γ. It follows from $H(x) = 0$ on I that $H^{(i)}(a_1) = 0$ for any $i \geq 0$. Thus $H(x) = 0$ on $(0, 2a_1)$. Now take $a_2 = \frac{3}{2}a_1$ and repeat the above analysis with a_1 replaced by a_2, resulting in $H(x) = 0$ on $(0, 2a_2) = (0, 3a_1)$. Then take $a_3 = \frac{3}{2}a_2$ and keep repeating the process, and one obtains $H(x) = 0$ on $(0, \infty)$ since $a_1 > 0$. Let γ_0 be the smallest power of all power functions that appear in $\{g_i(x)\}_{i=1}^k$, $\{h_i(x)\}_{i=1}^k$, and define $\tilde{H}(x) = x^{-\gamma_0}H(x)$. Then $\tilde{H}(x) = 0$ on $(0, \infty)$. Then

$$0 = \lim_{x\to\infty} e^{-b_k x}\tilde{H}(x) = \lim_{x\to\infty} (x^{-\gamma_0}h_k(x) + x^{-\gamma_0}g_k(x)\ln(x)),$$

which happens only when $x^{-\gamma_0}h_k(x) \equiv 0$ and $x^{-\gamma_0}g_k(x) \equiv 0$. That is, when $x \neq 0$, $h_k(x) \equiv 0$ and $g_k(x) \equiv 0$. Proceed in the same manner to show when $x \neq 0$, $h_i(x) \equiv 0$ and $g_i(x) \equiv 0$ for i from $k-1$ to 1.

□

2.11 Proofs in Section 2.5

This section contains all the proofs in Section 2.5 except that of Theorem 2.5.7 and Theorem 2.5.14. The proofs of Theorem 2.5.7 and Theorem 2.5.14 occupy the bulk of the chapter and will be presented in Section 2.12.

2.11.1 Proofs in Section 2.5.1

Proof of Lemma 2.5.4: In this proof we write n_1 and $\underline{\mathbf{N}}_1$ for $n_1(G_0)$ and $\underline{\mathbf{N}}_1(G_0)$ respectively. By Lemma 2.5.1 b), $n_1 = \underline{\mathbf{N}}_1 < \infty$. For each $N \geq 1$, there exists $R_N(G_0) > 0$ such that for any $G \in \mathcal{E}_{k_0}(\Theta)\backslash\{G_0\}$ and $W_1(G, G_0) < R_N(G_0)$

$$\frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)} \geq \frac{1}{2} \liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)}. \tag{2.70}$$

Take $c(G_0, N_0) = \min_{1\leq i\leq N_0} R_i(G_0) > 0$. Moreover, by the definition (2.27) for any $N \geq \underline{\mathbf{N}}_1$,

$$\liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)} \geq \inf_{N\geq\underline{\mathbf{N}}_1} \liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G\in\mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)} > 0.$$

Combining the last two displays completes the proof with

$$C(G_0) = \frac{1}{2} \inf_{N \geq \underline{\mathbf{N}}_1} \liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_N(G, G_0)}.$$

□

Proof of Lemma 2.5.5: In this proof we write n_1 for $n_1(G_0)$. By the definition of n_1, for any $N \geq n_1$

$$\liminf_{\substack{G \xrightarrow{W_1} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta_1)}} \frac{V(P_{G,N}, P_{G_0,N})}{D_1(G, G_0)} > 0. \tag{2.71}$$

By Lemma 2.3.2 b) one may replace the $D_1(G, G_0)$ in the preceding display by $W_1(G, G_0)$. Fix $N_1 = n_1 \vee n_0$. Then there exists $R > 0$ depending on G_0 such that

$$\inf_{G \in B_{W_1}(G_0, R) \setminus \{G_0\}} \frac{V(P_{G,N_1}, P_{G_0,N_1})}{W_1(G, G_0)} > 0, \tag{2.72}$$

where $B_{W_1}(G_0, R)$ is the open ball in metric space $(\bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1), W_1)$ with center at G_0 and radius R. Here we used the fact that any sufficiently small open ball in $(\bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1), W_1)$ with center in $\mathcal{E}_{k_0}(\Theta_1)$ is in $\mathcal{E}_{k_0}(\Theta_1)$.

Notice that $\bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1)$ is compact under the W_1 metric if Θ_1 is compact. By the assumption that the map $\theta \mapsto P_\theta$ is continuous and by Lemma 2.11.2 and the triangle inequality of total variation distance, $V(P_{G,N}, P_{G_0,N})$ with domain $(\bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1), W_1)$ is a continuous function of G for each N. Then $G \mapsto \frac{V(P_{G,N}, P_{G_0,N})}{W_1(G,G_0)}$ is a continuous map for each N. Moreover $\frac{V(P_{G,N}, P_{G_0,N})}{W_1(G,G_0)}$ is positive on the compact set $\bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1) \setminus B_{W_1}(G_0, R)$ provided $N \geq n_0$. As a result for each $N \geq n_0$

$$\min_{G \in \bigcup_{k=1}^{k_0} \mathcal{E}_k(\Theta_1) \setminus B_{W_1}(G_0, R)} \frac{V(P_{G,N}, P_{G_0,N})}{W_1(G, G_0)} > 0.$$

Combining the last display with $N_1 = n_1 \vee n_0$ and (2.72) yields

$$V(P_{G,N_1}, P_{G,N_1}) \geq C(G_0, \Theta_1) W_1(G, G_0), \tag{2.73}$$

where $C(G_0, \Theta_1)$ is a constant depending on G_0 and Θ_1. Observe $V(P_{G,N}, P_{G_0,N})$ increases with N, the proof is then complete. □

Proof of Lemma 2.5.6: It's easy to see when θ is a sufficiently small neighborhood of θ_i^0,

$$(2\|(J_g(\theta_i^0))^{-1}\|_2)^{-1} \|\theta - \theta_i^0\|_2 \leq \|g(\theta) - g(\theta_i^0)\|_2 \leq 2\|J_g(\theta_i^0)\|_2 \|\theta - \theta_i^0\|_2.$$

Then when G is in a small neighborhood of G_0 under W_1

$$(2\max_{1\le i\le k_0}\|(J_g(\theta_i^0))^{-1}\|_2+1)^{-1}D_N(G,G_0)\le D_N(G^\eta,G_0^\eta)\le(2\max_{1\le i\le k_0}\|J_g(\theta_i^0)\|_2+1)D_N(G,G_0).$$

Moreover $V(\tilde{P}_{G^\eta,N},\tilde{P}_{G_0^\eta,N})=V(P_{G,N},P_{G_0,N})$. Denote the left side and right side of (2.29) respectively by L and R. Then $L\le C(G_0)R$ and $L\ge c(G_0)R$ with

$$C(G_0)=2\max_{1\le i\le k_0}\|(J_g(\theta_i^0))^{-1}\|_2+1,\quad c(G_0)=(2\max_{1\le i\le k_0}\|J_g(\theta_i^0)\|_2+1)^{-1}.$$

The other equation in the statement follows similarly. □

The rest of this subsection contains auxiliary lemmas required in the previous proofs.

Lemma 2.11.1 (Lack of first-order identifiability)**.** *Fix $G_0=\sum_{i=1}^{k_0}p_i^0\delta_{\theta_i^0}\in\mathcal{E}_{k_0}(\Theta)$. Suppose*

$$\sum_{i=1}^{k_0}b_iP_{\theta_i^0}=0,\quad \sum_{i=1}^{k_0}b_i=0.$$

has a nonzero solution $(b_1,\ldots,b_{k_0})$. Here the 0 in the first equation is the zero measure on $\mathfrak{X}$. Then

$$\liminf_{\substack{G\xrightarrow{W_1}G_0\\ G\in\mathcal{E}_{k_0}(\Theta)}}\frac{V(P_G,P_{G_0})}{D_1(G,G_0)}=0. \tag{2.74}$$

Proof of Lemma 2.11.1: Construct $G_\ell=\sum_{i=1}^{k_0}p_i^\ell\delta_{\theta_i^0}$ with $p_i^\ell=p_i^0+b_i/\ell$ for $i\in[k_0]$. Then for large ℓ, $p_i^\ell\in(0,1)$ and $\sum_{i=1}^{k_0}p_i^\ell=1$. Then for large ℓ, $G_\ell\in\mathcal{E}_{k_0}(\Theta)$ and $G_\ell\xrightarrow{W_1}G_0$. Then the proof is complete by for large ℓ

$$V(P_{G_\ell},P_{G_0})=\sup_{A\in\mathcal{A}}|P_{G_\ell}(A)-P_{G_0}(A)|=\sup_{A\in\mathcal{A}}|1/\ell\sum_{i=1}^{k_0}b_iP_{\theta_i^0}(A)|=0.$$

and

$$D_1(G_\ell,G_0)=\frac{1}{\ell}\sum_{i=1}^{k_0}|b_i|\neq 0.$$

□

Lemma 2.11.2. *For any $G=\sum_{i=1}^{k_0}p_i\delta_{\theta_i}$ and $G'=\sum_{i=1}^{k_0}p_i'\delta_{\theta_i'}$,*

$$V(P_{G,N},P_{G',N})\le\begin{cases}\min_\tau\left(\sqrt{N}\max_{1\le i\le k_0}h\left(P_{\theta_i},P_{\theta'_{\tau(i)}}\right)+\frac{1}{2}\sum_{i=1}^{k_0}\left|p_i-p'_{\tau(i)}\right|\right), & N\ge 2,\\ \min_\tau\left(\max_{1\le i\le k_0}V\left(P_{\theta_i},P_{\theta'_{\tau(i)}}\right)+\frac{1}{2}\sum_{i=1}^{k_0}\left|p_i-p'_{\tau(i)}\right|\right), & N=1.\end{cases}$$

where the minimum is taken over all τ in the permutation group S_{k_0}.

Proof: The proof is similar as that of Lemma 2.7.2. □

Lemma 2.11.3. *Suppose the same conditions in Corollary 2.4.6 hold. Then for any $a \in \mathbb{R}^q$, for each $i \in [k_0]$, and for any $0 < \Delta \leq \gamma(\theta_i^0, a)$,*

$$\left| \frac{\prod_{j=1}^N f(x_j|\theta_i^0 + a\Delta) - \prod_{j=1}^N f(x_j|\theta_i^0)}{\Delta} \right| \leq \tilde{f}_\Delta(\bar{x}|\theta_i^0, a, N), \quad \bigotimes^N \mu\text{−}a.e.\, \bar{x} = (x_1, \ldots, x_N) \in \mathfrak{X}^N$$

where $\tilde{f}_\Delta(\bar{x}|\theta_i^0, a, N)$ satisfies

$$\lim_{\Delta \to 0^+} \int_{\mathfrak{X}^N} \tilde{f}_\Delta(\bar{x}|\theta_i^0, a, N) d \bigotimes^N \mu = \int_{\mathfrak{X}^N} \lim_{\Delta \to 0^+} \tilde{f}_\Delta(\bar{x}|\theta_i^0, a, N) d \bigotimes^N \mu.$$

Proof: By decomposing the difference as a telescoping sum,

$$\begin{aligned} &\left| \frac{\prod_{j=1}^N f(x_j|\theta_i^0 + a\Delta) - \prod_{j=1}^N f(x_j|\theta_i^0)}{\Delta} \right| \\ \leq &\sum_{\ell=1}^N \left(\prod_{j=1}^{\ell-1} f(x_j|\theta_i^0 + a\Delta) \right) \left| \frac{f(x_\ell|\theta_i^0 + a\Delta) - f(x_\ell|\theta_i^0)}{\Delta} \right| \left(\prod_{j=\ell+1}^N f(x_j|\theta_i^0) \right). \end{aligned}$$

Then the upper bound in the preceding display is upper bounded by

$$\tilde{f}_\Delta(\bar{x}|\theta_i^0, a, N) := \sum_{\ell=1}^N \left(\prod_{j=1}^{\ell-1} f(x_j|\theta_i^0 + a\Delta) \right) \bar{f}_\Delta(x_\ell|\theta_i^0, a) \left(\prod_{j=\ell+1}^N f(x_j|\theta_i^0) \right), \quad \bigotimes^N \mu\text{−}a.e.\, \bar{x} \in \mathfrak{X}^N.$$

For clean we write $\tilde{f}_\Delta(\bar{x}|\theta_i^0, a)$ for $\tilde{f}_\Delta(\bar{x}|\theta_i^0, a, N)$ in the rest of the proof. Notice $\tilde{f}_\Delta(\bar{x}|\theta_i^0, a)$ satisfies

$$\int_{\mathfrak{X}^N} \tilde{f}_\Delta(\bar{x}|\theta_i^0, a) d \bigotimes^N \mu = \sum_{\ell=1}^N \int_{\mathfrak{X}} \bar{f}_\Delta(x_\ell|\theta_i^0, a) d\mu = N \int_{\mathfrak{X}} \bar{f}_\Delta(x|\theta_i^0, a) d\mu \to N \int_{\mathfrak{X}} \lim_{\Delta \to 0^+} \bar{f}_\Delta(x|\theta_i^0, a) d\mu.$$

Moreover,

$$\lim_{\Delta \to 0^+} \tilde{f}_\Delta(\bar{x}|\theta_i^0, a) = \sum_{\ell=1}^N \left(\prod_{j=1}^{\ell-1} f(x_j|\theta_i^0) \right) \lim_{\Delta \to 0^+} \bar{f}_\Delta(x_\ell|\theta_i^0, a) \left(\prod_{j=\ell+1}^N f(x_j|\theta_i^0) \right), \quad \bigotimes^N \mu\text{−}a.e.\, \bar{x} \in \mathfrak{X}^N$$

and thus

$$\int_{\mathfrak{X}^N} \lim_{\Delta\to 0^+} \tilde{f}_\Delta(\bar{x}|\theta_i^0, a) d \bigotimes^N \mu = \sum_{\ell=1}^N \int_{\mathfrak{X}} \lim_{\Delta\to 0^+} \bar{f}_\Delta(x_\ell|\theta_i^0, a) d\mu = N \int_{\mathfrak{X}} \lim_{\Delta\to 0^+} \bar{f}_\Delta(x|\theta_i^0, a) d\mu.$$

□

2.11.2 Proofs and additional examples in Section 2.5.2

Proof of Corollary 2.5.8: Consider $\tilde{f}(x|\eta) := f(x|\theta)$ be the same kernel but under the new parameter $\eta = \eta(\theta)$. Note $\{\tilde{f}(x|\eta)\}_{\eta\in\Xi}$ with $\Xi := \eta(\Theta)$ is the canonical parametrization of the same exponential family. Write $\eta_i^0 = \eta(\theta_i^0)$. The proof is then completed by applying Lemma 2.5.7 to $\tilde{f}(x|\eta)$ and then by applying Lemma 2.5.6.

□

Lemma 2.11.4. *a) Let $\eta_1, \eta_2, \ldots, \eta_{2k}$ be $2k$ distinct real numbers. Let $n \leq 2k-2$. Then the system of $(2k-1)$ linear equations of $(y_1, y_2, \ldots, y_{2k})$*

$$\sum_{i=1}^{2k} y_i \eta_i^j = 0 \quad \forall j \in [n] \cup \{0\} \tag{2.75}$$

has all the solutions given by

$$y_i = -\sum_{q=n+2}^{2k} y_q \prod_{\substack{\ell\neq i \\ \ell=1}}^{n+1} \frac{(\eta_q - \eta_\ell)}{(\eta_i - \eta_\ell)} \quad \forall i \in [n+1] \tag{2.76}$$

for any $y_{n+2}, \ldots, y_{2k} \in \mathbb{R}$.

b) For any $0 < \eta_{k+1} < \eta_{k+2} < \ldots < \eta_{2k}$ and for any positive $y_{k+1}, y_{k+2}, \ldots, y_{2k}$, there exists infinitely many $\eta_1, \eta_2, \ldots, \eta_k$ satisfying

$$\eta_{k+i-1} < \eta_i < \eta_{k+i}, \quad \text{for } 2 \leq i \leq k, \text{ and } 0 < \eta_1 < \eta_{k+1} \text{ and}$$
$$y_i = -y_{2k} \prod_{\substack{\ell\neq i \\ \ell=1}}^{2k-1} \frac{(\eta_{2k} - \eta_\ell)}{(\eta_i - \eta_\ell)} \quad \forall k+1 \leq i \leq 2k-1.$$

c) For any $0 < \eta_{k+1} < \eta_{k+2} < \ldots < \eta_{2k}$ and for any positive $y_{k+1}, y_{k+2}, \ldots, y_{2k}$, the system of

equations of $(y_1, \ldots, y_k, \eta_1, \ldots, \eta_k)$

$$\sum_{i=1}^{2k} y_i \eta_i^j = 0 \quad \forall j \in [2k-2] \cup \{0\}$$

$$y_i < 0 \quad \forall i \in [k] \tag{2.77}$$

$$\eta_1 \in (0, \eta_{k+1}),\ \eta_i \in (\eta_{k+i-1}, \eta_{k+i}) \quad \forall 2 \le i \le k \tag{2.78}$$

has infinitely many solutions.

d) If $P_{G,n} = P_{G',n}$ *for some positive integer* n*, then* $P_{G,m} = P_{G',m}$ *for any integer* $1 \le m \le n$*.*

e) Consider the kernel specified in Example 2.5.10. For any $G \in \mathcal{E}_k(\Theta)$ *and for any* $n \le 2k-2$*, there exists infinitely many* $G' \in \mathcal{E}_k(\Theta)$ *such that* $P_{G,n} = P_{G',n}$*. In particular, this shows* $n_0(G) \ge 2k-1$ *for any* $G \in \mathcal{E}_k(\Theta)$*.*

Proof: a) By Lagrange interpolation formula over the points $\eta_1, \eta_2, \ldots, \eta_{n+1}$,

$$x^j = \sum_{i=1}^{n+1} \eta_i^j \prod_{\substack{\ell \ne i \\ \ell=1}}^{n+1} \frac{(x-\eta_\ell)}{(\eta_i - \eta_\ell)}, \quad \forall j \in [n] \cup \{0\},\ \forall x \in \mathbb{R}.$$

In particular, for any $n+2 \le q \le 2k$,

$$\eta_q^j = \sum_{i=1}^{n+1} \eta_i^j \prod_{\substack{\ell \ne i \\ \ell=1}}^{n+1} \frac{(\eta_q - \eta_\ell)}{(\eta_i - \eta_\ell)}, \quad \forall j \in [n] \cup \{0\}.$$

Plugging the above identity into (2.75), it is clear that the y_i specified in (2.76) are solutions of (2.75). Notice that the coefficient matrix of (2.75) is $A = (\eta_i^j)_{j \in [n] \cup \{0\}, i \in [2k]} \in \mathbb{R}^{(n+1)\times(2k)}$ has rank $n+1$ since the submatrix consisting the first $n+1$ columns form a non-singular Vandermonde matrix. Thus all the solutions of (2.75) form a subspace of $\mathbb{R}^{2k}$ of dimension $2k-(n+1)$, which implies (2.76) are all the solutions.

b) Let $a > 0$. Consider a polynomial $g(x)$ such that $g(0) = (-1)^{k+1}a$, $g(\eta_{2k}) = -\frac{1}{y_{2k}}$, and for $k+1 \le i \le 2k-1$, $g(\eta_i) = \frac{1}{y_i} \prod_{\substack{\ell \ne i \\ \ell=k+1}}^{2k-1} \frac{(\eta_{2k}-\eta_\ell)}{(\eta_i - \eta_\ell)}$. Then this $k+1$ points determines uniquely a polynomial $g(x)$ with degree at most k. By our construction, $g(x)$ satisfies

$$y_i g(\eta_i) = -y_{2k} g(\eta_{2k}) \prod_{\substack{\ell \ne i \\ \ell=k+1}}^{2k-1} \frac{(\eta_{2k} - \eta_\ell)}{(\eta_i - \eta_\ell)}, \quad \forall\ k+1 \le i \le 2k-1 \tag{2.79}$$

Moreover, noticing that $g(\eta_i) > 0$ for i odd integer between $k+1$ and $2k$, and $g(\eta_i) < 0$ for i even integer between $k+1$ and $2k$. Then there must exist $\eta_1 \in (0, \eta_{k+1})$ and $\eta_i \in (\eta_{k+i-1}, \eta_{k+i})$ for $2 \leq i \leq k$ such that $g(\eta_i) = 0$. Then $g(x) = b\prod_{i=1}^{k}(x - \eta_i)$ where $b < 0, \eta_1, \eta_2, \ldots, \eta_k$ are constants that depend on $a, \eta_{k+1}, \ldots, \eta_{2k}, y_{k+1}, \ldots, y_{2k}$. Plug $g(x) = b\prod_{i=1}^{k}(x - \eta_i)$ into (2.79) shows that $(\eta_1, \eta_2, \ldots, \eta_k)$ is a solution for the system of equations in the statement. By changing value of a, we get infinitely many solutions.

c) First, we apply part a) with $n = 2k-2$: for any $2k$ distinct real numbers $\eta_1, \ldots, \eta_{2k}$, the system of linear equations of $(x_1, \ldots, x_{2k})$

$$\sum_{i=1}^{2k} x_i \eta_i^j = 0 \quad \forall j \in [2k-2] \cup \{0\}$$

has a solution

$$x_i = -y_{2k} \prod_{\substack{\ell \neq i \\ \ell=1}}^{2k-1} \frac{(\eta_{2k} - \eta_\ell)}{(\eta_i - \eta_\ell)} \quad \forall i \in [2k-1],$$

where we have specified $x_{2k} = y_{2k}$.

Next, for the $\eta_{k+1}, \ldots, \eta_{2k}$ given in the lemma's statement, by part b) we can choose $\eta_1, \ldots, \eta_k$ that satisfy the requirements there. Accordingly, $x_i = y_i$ for $k+1 \leq i \leq 2k$. Moreover, it follows from the ranking of $\{\eta_i\}_{i=1}^{2k}$ that $x_i < 0$ for any $i \in [k]$. Thus $(x_1, \ldots, x_k, \eta_1, \ldots, \eta_k)$ is a solution of the system of equations in the statement. The infinite many solutions conclusion follows since there are infinitely many $(\eta_1, \ldots, \eta_k)$ by part b).

d) $P_{G,n-1} = P_{G',n-1}$ follows immediately from for any $A \in \mathcal{A}^{n-1}$, the product sigma-algebra on $\mathfrak{X}^{n-1}$,

$$P_{G,n-1}(A) = P_{G,n}(A \times \mathfrak{X}) = P_{G',n}(A \times \mathfrak{X}) = P_{G',n-1}(A).$$

Repeating this procedure inductively and the conclusion follows.

e) By part d) it suffices to prove that $n = 2k-2$. Write $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ with $\theta_1 < \theta_2 < \ldots < \theta_k$. Consider any $G' = \sum_{i=1}^{k} p_i' \delta_{\theta_i'} \in \mathcal{E}_k(\Theta)$ with $\theta_1' < \theta_2' < \ldots < \theta_k'$ such that $P_{G,n} = P_{G',n}$. $P_{G,n} = P_{G',n}$ for $n = 2k-2$ is

$$\sum_{i=1}^{k} p_i'(\theta_i')^j(1-\theta_i')^{2k-2-j} = \sum_{i=1}^{k} p_i(\theta_i)^j(1-\theta_i)^{2k-2-j} \quad \forall j = 0, 1, \cdots, 2k-2. \tag{2.80}$$

$$0 < \theta_1' < \ldots < \theta_k' < 1, p_i' > 0, \ \forall i \in [k] \tag{2.81}$$

Note the system of equations (2.80) automatically implies $\sum_{i=1}^{k} p_i' = \sum_{i=1}^{k} p_i = 1$. Let $y_i =$

$-p_i'(1-\theta_i')^{2k-2}$, $\eta_i = \theta_i'/(1-\theta_i')$ for $i \in [k]$ and let $y_{k+i} = p_i(1-\theta_i)^{2k-2}$, $\eta_{k+i} = \theta_i/(1-\theta_i)$. Then $\eta_{k+1} < \eta_{k+2} < \ldots < \eta_{2k}$ and $y_i > 0$ for $k+1 \leq i \leq 2k$. Then $(p_1', \ldots, p_k', \theta_1', \ldots, \theta_k')$ is a solution of (2.80), (2.81) if and only if the corresponding $(y_1, \ldots, y_k, \eta_1, \ldots, \eta_k)$ is the solution of

$$\sum_{i=1}^{2k} y_i \eta_i^j = 0, \quad \forall j \in [2k-2] \cup \{0\}.$$
$$0 < \eta_1 < \ldots < \eta_k,\ y_i < 0,\ \forall i \in [k].$$

By part c), the system of equations in last display has infinitely many solutions additionally satisfying (2.78). For each such solution, the corresponding $(p_1', \ldots, p_k', \theta_1', \ldots, \theta_k')$ is a solution of system of equations (2.80) (2.81) additionally satisfying $0 < \theta_1' < \theta_1$ and $\theta_{i-1} < \theta_i' < \theta_i$ for $2 \leq i \leq k$. By the comments after (2.80),(2.81) we also have $\sum_{i=1}^k p_i' = \sum_{i=1}^k p_i = 1$. Thus, such $(p_1', \ldots, p_k', \theta_1', \ldots, \theta_k')$ gives $G' \in \mathcal{E}_k(\Theta)$ such that $P_{G',2k-2} = P_{G,2k-2}$. The existence of infinitely many such G' follows from the existence of infinitely many solutions $(y_1, \ldots, y_k, \eta_1, \ldots, \eta_k)$ by part c).

□

Proof of Lemma 2.5.11: a) It's obvious that $q^{(1)}(x,y), q^{(2)}(x,y)$ are multivariate polynomials and that

$$q^{(1)}(y,y) = \lim_{x \to y} q^{(1)}(x,y) = f'(y),$$
$$q^{(2)}(y,y) = \lim_{x \to y} q^{(2)}(x,y) = f''(y).$$

That means $q^{(1)}(x,y) - f'(y)$ has factor $x-y$ and thus $\bar{q}^{(2)}(x,y)$ is a multivariate polynomial and

$$\bar{q}^{(2)}(y,y) = \lim_{x \to y} \frac{q^{(1)}(x,y) - f'(y)}{x-y}$$
$$= \lim_{x \to y} \frac{f(x) - f(y) - f'(y)(x-y)}{(x-y)^2} = \frac{1}{2} f''(y) = \frac{1}{2} q^{(2)}(y,y).$$

Then $\bar{q}^{(2)}(x,y) - \frac{1}{2} q^{(2)}(x,y)$ has factor $x-y$ and thus $\bar{q}^{(3)}(x,y)$ is a multivariate polynomial.

b) Write $A^{(k)}$ for $A^{(k)}(x_1, \ldots, x_k)$ in this proof. Denote $\underline{A} \in \mathbb{R}^{(2k-2)\times(2k)}$ the bottom $(2k-2) \times 2k$ matrix of $A^{(k)}$. Let $q_j^{(1)}(x,y)$, $q_j^{(2)}(x,y)$, $\bar{q}_j^{(2)}(x,y)$ and $\bar{q}_j^{(3)}(x,y)$ be defined in part a) with f replace by f_j. Then by subtracting the third row from the first row, the fourth row from the

second row and then factor the common factor $(x_1 - x_2)$ out of the resulting first two rows

$$\begin{aligned}
\det(A^{(k)}) =&(x_1-x_2)^2\det\begin{pmatrix} q_1^{(1)}(x_1,x_2), & \ldots, & q_{2k}^{(1)}(x_1,x_2) \\ q_1^{(2)}(x_1,x_2), & \ldots, & q_{2k}^{(2)}(x_1,x_2) \\ & \underline{\mathrm{A}} & \end{pmatrix} \\
=&(x_1-x_2)^3\det\begin{pmatrix} \bar{q}_1^{(2)}(x_1,x_2), & \ldots, & \bar{q}_{2k}^{(2)}(x_1,x_2) \\ q_1^{(2)}(x_1,x_2), & \ldots, & q_{2k}^{(2)}(x_1,x_2) \\ & \underline{\mathrm{A}} & \end{pmatrix} \\
=&(x_1-x_2)^4\det\begin{pmatrix} \bar{q}_1^{(3)}(x_1,x_2), & \ldots, & \bar{q}_{2k}^{(3)}(x_1,x_2) \\ q_1^{(2)}(x_1,x_2), & \ldots, & q_{2k}^{(2)}(x_1,x_2) \\ & \underline{\mathrm{A}} & \end{pmatrix}
\end{aligned}$$

where the second equality follows by subtracting the fourth row from first row and then factor the common factor $(x_1 - x_2)$ out of the resulting row. The last step of the preceding display follows by subtracting 1/2 times the second row and then extract the common factor $(x_1 - x_2)$ out of the resulting row. Thus $(x_1 - x_2)^4$ is a factor of $\det(A^{(k)})$, which is a multivariate polynomial in $x_1, \ldots, x_k$. By symmetry, $\prod_{1\le\alpha<\beta\le k}(x_\alpha - x_\beta)^4$ is a factor of $\det(A^{(k)})$.

c) We prove $\det(A^{(k)}(x_1, \ldots, x_k)) = \prod_{1\le\alpha<\beta\le k}(x_\alpha - x_\beta)^4$ by induction. It's easy to verify the statement holds when $k = 1$. Suppose the statement for k holds. By b),

$$\det(A^{(k+1)}(x_1,\ldots,x_{k+1})) = g_{k+1}(x_1,\ldots,x_{k+1}) \prod_{1\le\alpha<\beta\le k+1}(x_\alpha - x_\beta)^4$$

for some multivariate polynomial g_{k+1}. Since $f_j(x)$ has degree $j-1$, $f_j'(x)$ has degree $j-2$, and hence by Leibniz formula of determinant
$\det(A^{(k+1)}(x_1, \ldots, x_k, x_{k+1}))$ has degree no more than $2k+2k = 4k$ for any x_α for $\alpha \in [k+1]$. Moreover, in $\prod_{1\le\alpha<\beta\le k+1}(x_\alpha - x_\beta)^4$ the degree of x_α is $4k$ and the corresponding term is x_α^{4k}, which implies in $g_{k+1}(x_1, \ldots, x_{k+1})$ the degree of x_α is no more than 0 for any $\alpha \in [k+1]$. As a result, $g_{k+1}(x_1, \ldots, x_{k+1}) = q_{k+1}$ is a constant. Thus

$$\det(A^{(k+1)}(x_1,\ldots,x_k,0)) = q_{k+1}\left(\prod_{1\le\alpha<\beta\le k}(x_\alpha - x_\beta)^4\right)\prod_{\alpha=1}^{k} x_\alpha^4, \tag{2.82}$$

On the other hand,

$$\det(A^{(k+1)}(x_1,\ldots,x_k,0))$$

$$
\begin{aligned}
&= \det \begin{pmatrix}
f_1(x_1|k+1), & f_2(x_1|k+1), & \dots, & f_{2(k+1)}(x_1|k+1) \\
f_1'(x_1|k+1), & f_2'(x_1|k+1), & \dots, & f_{2(k+1)}'(x_1|k+1) \\
\vdots & \vdots & \vdots & \\
f_1(x_k|k+1), & f_2(x_k|k+1), & \dots, & f_{2(k+1)}(x_k|k+1) \\
f_1'(x_k|k+1), & f_2'(x_k|k+1), & \dots, & f_{2(k+1)}'(x_k|k+1) \\
1, & 0, & \dots, & 0 \\
0, & 1, & 0, \dots, & 0
\end{pmatrix} \\
&= \det \begin{pmatrix}
f_3(x_1|k+1), & f_3(x_1|k+1), & \dots, & f_{2(k+1)}(x_1|k+1) \\
f_3'(x_1|k+1), & f_3'(x_1|k+1), & \dots, & f_{2(k+1)}'(x_1|k+1) \\
\vdots & \vdots & \vdots & \\
f_3(x_k|k+1), & f_3(x_k|k+1), & \dots, & f_{2(k+1)}(x_k|k+1) \\
f_3'(x_k|k+1), & f_3'(x_k|k+1), & \dots, & f_{2(k+1)}'(x_k|k+1)
\end{pmatrix}
\end{aligned} \tag{2.83}
$$

where the second equality follows by Laplace expansion along the last row. Observing $f_j(x) = x^2 f_{j-2}(x)$ and $f_j'(x) = x^2 f_{j-2}'(x) + 2x f_{j-2}(x)$, plug these two equations into (2.83) and simplify the resulting determinant,

$$
\det(A^{(k+1)}(x_1, \dots, x_k, 0)) = \det(A^{(k)}(x_1, \dots, x_k)) \prod_{\alpha=1}^{k} x_\alpha^4. \tag{2.84}
$$

Compare (2.84) to (2.82), together with the induction assumption that statement for k holds,

$$
q_{k+1} = 1.
$$

That is, we proved the statement for $k+1$.

d) We prove $\det(A^{(k)}(x_1, \dots, x_k)) = \prod_{1 \le \alpha < \beta \le k} (x_\alpha - x_\beta)^4$ by induction. Write $f_j(x|k)$ for $f_j(x)$ in the following induction to emphasize its dependence on k. It's easy to verify the case holds when $k = 1$. Suppose the statement for k holds. By b), $\det(A^{(k+1)}(x_1, \dots, x_{k+1})) = g_{k+1}(x_1, \dots, x_{k+1}) \prod_{1 \le \alpha < \beta \le k+1} (x_\alpha - x_\beta)^4$ for some multivariate polynomial g_{k+1}. Since $f_j(x|k+1)$ has degree $n = 2(k+1) - 1$, $f_j'(x|k+1)$ has degree $2k$, and hence by Leibniz formula of determinant $\det(A^{(k+1)}(x_1, \dots, x_k, x_{k+1}))$ has degree no more than $2k + (2k+1) = 4k+1$ for any x_α for $\alpha \in [k+1]$. Moreover, in $\prod_{1 \le \alpha < \beta \le k+1} (x_\alpha - x_\beta)^4$ the degree of x_α is $4k$, which implies in $g_{k+1}(x_1, \dots, x_{k+1})$ the degree of x_α is no more than 1. As a result, it's eligible to write $g_{k+1}(x_1, \dots, x_{k+1}) = h_1(x_1, \dots, x_k) x_{k+1} + h_2(x_1, \dots, x_k)$ where h_1, h_2 are

multivariate polynomials of $x_1, \ldots, x_k$. Thus

$$\det(A^{(k+1)}(x_1, \ldots, x_k, 0)) = h_2(x_1, \ldots, x_k) \left(\prod_{1 \le \alpha < \beta \le k} (x_\alpha - x_\beta)^4 \right) \prod_{\alpha=1}^{k} x_\alpha^4, \tag{2.85}$$

$$\begin{aligned} &\det(A^{(k+1)}(x_1, \ldots, x_k, 1)) \\ =&(h_1(x_1, \ldots, x_k) + h_2(x_1, \ldots, x_k)) \left(\prod_{1 \le \alpha < \beta \le k} (x_\alpha - x_\beta)^4 \right) \prod_{\alpha=1}^{k} (x_\alpha - 1)^4. \end{aligned} \tag{2.86}$$

On the other hand,

$$\begin{aligned} &\det(A^{(k+1)}(x_1, \ldots, x_k, 0)) \\ =&\det \begin{pmatrix} f_1(x_1|k+1), & f_2(x_1|k+1), & \ldots, & f_{2(k+1)}(x_1|k+1) \\ f_1'(x_1|k+1), & f_2'(x_1|k+1), & \ldots, & f_{2(k+1)}'(x_1|k+1) \\ \vdots & \vdots & \vdots & \\ f_1(x_k|k+1), & f_2(x_k|k+1), & \ldots, & f_{2(k+1)}(x_k|k+1) \\ f_1'(x_k|k+1), & f_2'(x_k|k+1), & \ldots, & f_{2(k+1)}'(x_k|k+1) \\ 1, & 0, & \ldots, & 0 \\ -(2(k+1)-1), & 1, & 0, \ldots, & 0 \end{pmatrix} \\ =&\det \begin{pmatrix} f_3(x_1|k+1), & f_3(x_1|k+1), & \ldots, & f_{2(k+1)}(x_1|k+1) \\ f_3'(x_1|k+1), & f_3'(x_1|k+1), & \ldots, & f_{2(k+1)}'(x_1|k+1) \\ \vdots & \vdots & \vdots & \\ f_3(x_k|k+1), & f_3(x_k|k+1), & \ldots, & f_{2(k+1)}(x_k|k+1) \\ f_3'(x_k|k+1), & f_3'(x_k|k+1), & \ldots, & f_{2(k+1)}'(x_k|k+1) \end{pmatrix} \end{aligned} \tag{2.87}$$

where the second equality follows by Laplace expansion along the last row. Observing $f_j(x|k+1) = x^2 f_{j-2}(x|k)$ and $f_j'(x|k+1) = x^2 f_{j-2}'(x|k) + 2x f_{j-2}(x|k)$, plug these two equations into (2.87) and simplify the resulting determinant,

$$\det(A^{(k+1)}(x_1, \ldots, x_k, 0)) = \det(A^{(k)}(x_1, \ldots, x_k)) \prod_{\alpha=1}^{k} x_\alpha^4. \tag{2.88}$$

Analogous argument produces

$$\det(A^{(k+1)}(x_1, \ldots, x_k, 1)) = \det(A^{(k)}(x_1, \ldots, x_k)) \prod_{\alpha=1}^{k} (1 - x_\alpha)^4. \tag{2.89}$$

Compare (2.88) to (2.85), together with the induction assumption that statement for k holds,

$$h_2(x_1, \ldots, x_k) = 1, \quad \forall x_1, \ldots, x_k.$$

Compare (2.89) to (2.86), together with the induction assumption that statement for k holds and the preceding display,

$$h_1(x_1, \ldots, x_k) = 0, \quad \forall x_1, \ldots, x_k.$$

That is, $g_{k+1}(x_1, \ldots, x_{k+1}) = 1$ for any $x_1, \ldots, x_{k+1}$.

□

2.12 Proof of inverse bounds for mixtures of product distributions

For an overview of our proof techniques, please refer to Section 2.2. The proofs of both Theorem 2.5.7 and Theorem 2.5.14 follow the same structure. The reader should read the former first before attempting the latter, which is considerably more technical and lengthy.

2.12.1 Proof of Theorem 2.5.7

Proof of Theorem 2.5.7:

Step 1 (Proof by contradiction with subsequences)

Suppose (2.24) is not true. Then $\exists\{N_\ell\}_{\ell=1}^{\infty}$ subsequence of natural numbers tending to infinity such that

$$\lim_{r\to 0} \inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}} \frac{V(P_{G,N_\ell}, P_{H,N_\ell})}{D_{N_\ell}(G,H)} \to 0 \quad \text{as } N_\ell \to \infty.$$

Then $\exists\{G_\ell\}_{\ell=1}^{\infty}, \{\Psi_\ell\}_{\ell=1}^{\infty} \subset \mathcal{E}_{k_0}(\Theta)$ such that

$$\begin{cases} G_\ell \neq H_\ell & \forall \ell \\ D_{N_\ell}(G_\ell, G_0) \to 0, D_{N_\ell}(H_\ell, G_0) \to 0 & \text{as } \ell \to \infty \\ \frac{V(P_{G_\ell,N_\ell}, P_{H_\ell,N_\ell})}{D_{N_\ell}(G_\ell,H_\ell)} \to 0 & \text{as } \ell \to \infty. \end{cases} \tag{2.90}$$

To see this, for each fixed ℓ, and thus fixed N_ℓ, $D_{N_\ell}(G, G_0) \to 0$ if and only if $W_1(G, G_0) \to 0$. Thus, there exists $G_\ell, H_\ell \in \mathcal{E}_{k_0}(\Theta)$ such that $G_\ell \neq H_\ell$, $D_{N_\ell}(G_\ell, G_0) \leq \frac{1}{\ell}$, $D_{N_\ell}(H_\ell, G_0) \leq \frac{1}{\ell}$ and

$$\frac{V(P_{G_\ell,N_\ell}, P_{H_\ell,N_\ell})}{D_{N_\ell}(G_\ell, H_\ell)} \leq \lim_{r\to 0} \inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}} \frac{V(P_{G,N_\ell}, P_{H,N_\ell})}{D_{N_\ell}(G,H)} + \frac{1}{\ell},$$

thereby ensuring that (2.90) hold.

Write $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$. We may relabel the atoms of G_ℓ and H_ℓ such that $G_\ell = \sum_{i=1}^{k_0} p_i^\ell \delta_{\theta_i^\ell}$, $H_\ell = \sum_{i=1}^{k_0} \pi_i^\ell \delta_{\eta_i^\ell}$ with $\theta_i^\ell, \eta_i^\ell \to \theta_i^0$ and $p_i^\ell, \pi_i^\ell \to p_i^0$ for any $i \in [k_0]$. By subsequence argument if necessary, we may require $\{G_\ell\}_{\ell=1}^\infty$, $\{H_\ell\}_{\ell=1}^\infty$ additionally satisfy:

$$\frac{\sqrt{N_\ell}\left(\theta_i^\ell - \eta_i^\ell\right)}{D_{N_\ell}(G_\ell, H_\ell)} \to a_i \in \mathbb{R}^q, \quad \frac{p_i^\ell - \pi_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \to b_i \in \mathbb{R}, \quad \forall 1 \le i \le k_0, \tag{2.91}$$

where the components of a_i are in $[-1, 1]$ and $\sum_{i=1}^{k_0} b_i = 0$. It also follows that at least one of a_i is not $\mathbf{0} \in \mathbb{R}^s$ or one of b_i is not 0. Let $\alpha \in \{1 \le i \le k_0 : a_i \ne \mathbf{0} \text{ or } b_i \ne 0\}$.

Step 2 (Change of measure by index α and application of CLT)

$P_{\theta,N}$ has density w.r.t. $\bigotimes^N \mu$ on $\mathfrak{X}^N$:

$$\bar{f}(\bar{x}|\theta, N) = \prod_{j=1}^N f(x_j|\theta) = \exp\left(\theta^T \left(\sum_{j=1}^N T(x_j)\right) - NA(\theta)\right) \prod_{j=1}^N h(x_j),$$

where any $\bar{x} \in \mathfrak{X}^N$ is partitioned into N blocks as $\bar{x} = (x_1, x_2, \ldots, x_N)$ with $x_i \in \mathfrak{X}$. Then

$$\begin{aligned}
&\frac{2V(P_{G_\ell,N_\ell}, P_{H_\ell,N_\ell})}{D_{N_\ell}(G_\ell, H_\ell)} \\
=&\int_{\mathfrak{X}^{N_\ell}} \left| \sum_{i=1}^{k_0} \frac{p_i^\ell \exp\left(\left\langle \theta_i^\ell, \sum_{j=1}^{N_\ell} T(x_j)\right\rangle - N_\ell A(\theta_i^\ell)\right) - \pi_i^\ell \exp\left(\left\langle \eta_i^\ell, \sum_{j=1}^{N_\ell} T(x_j)\right\rangle - N_\ell A(\eta_i^\ell)\right)}{D_{N_\ell}(G_\ell, H_\ell)} \right| \times \\
&\prod_{j=1}^{N_\ell} h(x_j) d \bigotimes^{N_\ell} \mu \\
=&\int_{\mathfrak{X}^{N_\ell}} \left| \sum_{i=1}^{k_0} \frac{p_i^\ell \exp\left(\left\langle \theta_i^\ell, \sum_{j=1}^{N_\ell} T(x_j)\right\rangle - N_\ell A(\theta_i^\ell)\right) - \pi_i^\ell \exp\left(\left\langle \eta_i^\ell, \sum_{j=1}^{N_\ell} T(x_j)\right\rangle - N_\ell A(\eta_i^\ell)\right)}{D_{N_\ell}(G_\ell, H_\ell) \exp\left(\left\langle \theta_\alpha^0, \sum_{j=1}^{N_\ell} T(x_j)\right\rangle - N_\ell A(\theta_\alpha^0)\right)} \right| \times \\
&\bar{f}(\bar{x}|\theta_\alpha^0, N_\ell) d \bigotimes^{N_\ell} \mu \\
=&\mathbb{E}_{\theta_\alpha^0} \left| F_\ell\left(\sum_{j=1}^{N_\ell} T(X_j)\right) \right|,
\end{aligned} \tag{2.92}$$

where X_j are i.i.d. random variables having densities $f(\cdot|\theta_\alpha^0)$, and

$$F_\ell(y) := \sum_{i=1}^{k_0} \frac{p_i^\ell \exp\left(\left\langle \theta_i^\ell, y\right\rangle - N_\ell A(\theta_i^\ell)\right) - \pi_i^\ell \exp\left(\left\langle \eta_i^\ell, y\right\rangle - N_\ell A(\eta_i^\ell)\right)}{D_{N_\ell}(G_\ell, H_\ell) \exp\left(\left\langle \theta_\alpha^0, y\right\rangle - N_\ell A(\theta_\alpha^0)\right)}.$$

Let $Z_\ell = \left(\sum_{j=1}^{N_\ell} T(X_j) - N_\ell \mathbb{E}_{\theta_\alpha^0} T(X_j)\right)/\sqrt{N_\ell}$. Then since $\theta_\alpha^0 \in \Theta^\circ$, the mean and covariance matrix of $T(X_j)$ are respectively $\nabla_\theta A(\theta_\alpha^0)$ and $\nabla_\theta^2 A(\theta_\alpha^0)$, the gradient and Hessian of $A(\theta)$ evaluated at θ_α^0. Then by central limit theorem, Z_ℓ converges in distribution to $Z \sim \mathcal{N}(\mathbf{0}, \nabla_\theta^2 A(\theta_\alpha^0))$. Moreover,

$$F_\ell\left(\sum_{j=1}^{N_\ell} T(X_j)\right) = F_\ell\left(\sqrt{N_\ell} Z_\ell + N_\ell \nabla_\theta A(\theta_\alpha^0)\right) := \Psi_\ell(Z_\ell), \tag{2.93}$$

where $\Psi_\ell(z) = F_\ell\left(\sqrt{N_\ell} z + N_\ell \nabla_\theta A(\theta_\alpha^0)\right)$.

Step 3 (Application of continuous mapping theorem) Define $\Psi(z) = p_\alpha^0 \langle a_\alpha, z\rangle + b_\alpha$. Suppose:

$$\Psi_\ell(z_\ell) \to \Psi(z) \text{ for any sequence } z_\ell \to z \in \mathbb{R}^q, \tag{2.94}$$

a property to be verified in the sequel, then by Generalized Continuous Mapping Theorem ([WVdV96] Theorem 1.11.1), $\Psi_\ell(Z_\ell)$ convergence in distribution to $\Psi(Z)$. Apply Theorem 25.11 in [Bil96],

$$\mathbb{E}|\Psi(Z)| \le \liminf_{\ell\to\infty} \mathbb{E}_{\theta_\alpha^0} |\Psi_\ell(Z_\ell)| = 0, \tag{2.95}$$

where the equality follows (2.90), (2.92) and (2.93). Since $\Psi(z)$ is a non-zero affine transform and the covariance matrix of Z is positive definite due to full rank property of exponential family, $\Psi(Z)$ is either a nondegenerate gaussian random variable or a non-zero constant, which contradicts with (2.95).

It remains in the proof to verify (2.94). Consider any sequence $z_\ell \to z$. Write

$$\Psi_\ell(z_\ell) = \sum_{i=1}^{k_0} I_i, \tag{2.96}$$

where

$$I_i := \frac{p_i^\ell \exp\left(g_\ell(\theta_i^\ell)\right) - \pi_i^\ell \exp\left(g_\ell(\eta_i^\ell)\right)}{D_{N_\ell}(G_\ell, H_\ell)\exp\left(g(\theta_\alpha^0)\right)},$$

with

$$g_\ell(\theta) = \left\langle \theta, \sqrt{N_\ell} z_\ell + N_\ell \nabla_\theta A(\theta_\alpha^0)\right\rangle - N_\ell A(\theta).$$

For any $i \in [k_0]$, by Taylor expansion of $A(\theta)$ at θ_i^0 and the fact that $A(\theta)$ is infinitely differentiable at $\theta_i^0 \in \Theta^\circ$, for large ℓ,

$$|A(\eta_i^\ell) - A(\theta_i^0) - \langle \nabla A(\theta_i^0), \eta_i^\ell - \theta_i^0\rangle| \le 2\|\nabla^2 A(\theta_i^0)\|_2 \|\eta_i^\ell - \theta_i^0\|_2^2,$$

which implies

$$\lim_{\ell\to\infty} N_\ell|A(\eta_i^\ell) - A(\theta_i^0) - \langle\nabla A(\theta_i^0), \eta_i^\ell - \theta_i^0\rangle| \le 2\|\nabla^2 A(\theta_i^0)\|_2 \lim_{\ell\to\infty} D^2_{N_\ell}(H_\ell, G_0) = 0 \quad (2.97)$$

where the equality follows from (2.90), and the inequality follows from that

$$D_{N_\ell}(H_\ell, G_0) = \sum_{i=1}^{k_0}(\sqrt{N_\ell}\|\eta_i^\ell - \theta_i^0\|_2 + |\pi_i^\ell - p_i^0|) \quad (2.98)$$

$$D_{N_\ell}(G_\ell, G_0) = \sum_{i=1}^{k_0}(\sqrt{N_\ell}\|\theta_i^\ell - \theta_i^0\|_2 + |p_i^\ell - p_i^0|) \quad (2.99)$$

for large ℓ. The same conclusion holds with η_i^ℓ replaced by θ_i^ℓ in the last two displays.

For $i \in [k_0]$, by Lemma 2.10.3 b) and the fact that $A(\theta)$ is infinitely differentiable at $\theta_i^0 \in \Theta^\circ$, for large ℓ

$$|A(\theta_i^\ell) - A(\eta_i^\ell) - \langle\nabla A(\theta_i^0), \theta_i^\ell - \eta_i^\ell\rangle| \le 2\|\nabla^2 A(\theta_i^0)\|_2\|\theta_i^\ell - \eta_i^\ell\|_2(\|\theta_i^\ell - \theta_i^0\|_2 + \|\eta_i^\ell - \eta_i^0\|_2),$$

which implies

$$\begin{aligned}
&\lim_{\ell\to\infty} \frac{N_\ell|A(\theta_i^\ell) - A(\eta_i^\ell) - \langle\nabla A(\theta_i^0), \theta_i^\ell - \eta_i^\ell\rangle|}{D_{N_\ell}(G_\ell, H_\ell)} \\
\le& 2\|\nabla^2 A(\theta_i^0)\|_2 \lim_{\ell\to\infty} \frac{\sqrt{N_\ell}\|\theta_i^\ell - \eta_i^\ell\|_2}{D_{N_\ell}(G_\ell, H_\ell)}(D_{N_\ell}(G_\ell, G_0) + D_{N_\ell}(H_\ell, G_0)) \\
=& 0
\end{aligned} \quad (2.100)$$

where the inequality follows from (2.98) and (2.99), and the equality follows from (2.90) and (2.91).

Case 1: Calculate $\lim_{\ell\to\infty} I_\alpha$.

When $\ell \to \infty$

$$g_\ell(\eta_\alpha^\ell) - g_\ell(\theta_\alpha^0) = \left\langle \eta_\alpha^\ell - \theta_\alpha^0, \sqrt{N_\ell} z_\ell \right\rangle - N_\ell \left(A(\eta_\alpha^\ell) - A(\theta_\alpha^0) - \left\langle \eta_\alpha^\ell - \theta_\alpha^0, \nabla_\theta A(\theta_\alpha^0)\right\rangle\right) \to 0 \quad (2.101)$$

by (2.90) and (2.97) with $i = \alpha$. Similarly, one has

$$\lim_{\ell\to\infty} \left(g_\ell(\theta_\alpha^\ell) - g_\ell(\theta_\alpha^0)\right) = 0 \quad (2.102)$$

Moreover when $\ell \to \infty$

$$\frac{g_\ell(\theta_\alpha^\ell) - g_\ell(\eta_\alpha^\ell)}{D_{N_\ell}(G_\ell, H_\ell)} = \frac{\left\langle \theta_\alpha^\ell - \eta_\alpha^\ell, \sqrt{N_\ell} z_\ell \right\rangle - N_\ell \left(A(\theta_\alpha^\ell) - A(\eta_\alpha^\ell) - \left\langle \theta_\alpha^\ell - \eta_\alpha^\ell, \nabla_\theta A(\theta_\alpha^0) \right\rangle \right)}{D_{N_\ell}(G_\ell, H_\ell)} \to \langle a_\alpha, z \rangle \tag{2.103}$$

by (2.91) and (2.100) with $i = \alpha$.

Thus

$$\begin{aligned}
&\lim_{\ell\to\infty} I_\alpha \\
=& \lim_{\ell\to\infty} \frac{p_\alpha^\ell \exp\left(g_\ell(\theta_\alpha^\ell) - g_\ell(\theta_\alpha^0)\right) - \pi_\alpha^\ell \exp\left(g_\ell(\eta_\alpha^\ell) - g_\ell(\theta_\alpha^0)\right)}{D_{N_\ell}(G_\ell, H_\ell)} \\
=& \lim_{\ell\to\infty} p_\alpha^\ell \frac{\exp\left(g_\ell(\theta_\alpha^\ell) - g_\ell(\theta_\alpha^0)\right) - \exp\left(g_\ell(\eta_\alpha^\ell) - g_\ell(\theta_\alpha^0)\right)}{D_{N_\ell}(G_\ell, H_\ell)} + \\
& \lim_{\ell\to\infty} \frac{p_\alpha^\ell - \pi_\alpha^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \exp\left(g_\ell(\eta_\alpha^\ell) - g_\ell(\theta_\alpha^0)\right) \\
\overset{(*)}{=}& p_\alpha^0 \lim_{\ell\to\infty} \frac{\exp(\xi_\ell)\left(g_\ell(\theta_\alpha^\ell) - g_\ell(\eta_\alpha^\ell)\right)}{D_{N_\ell}(G_\ell, H_\ell)} + \lim_{\ell\to\infty} \frac{p_\alpha^\ell - \pi_\alpha^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \exp\left(g_\ell(\eta_\alpha^\ell) - g_\ell(\theta_\alpha^0)\right) \\
\overset{(**)}{=}& p_\alpha^0 \lim_{\ell\to\infty} \frac{g_\ell(\theta_\alpha^\ell) - g_\ell(\eta_\alpha^\ell)}{D_{N_\ell}(G_\ell, H_\ell)} + \lim_{\ell\to\infty} \frac{p_\alpha^\ell - \pi_\alpha^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \\
\overset{(***)}{=}& p_\alpha^0 \langle a_\alpha, z \rangle + b_\alpha,
\end{aligned} \tag{2.104}$$

where step $(*)$ follows from mean value theorem with ξ_ℓ on the line segment between $g_\ell(\theta_\alpha^\ell) - g_\ell(\theta_\alpha^0)$ and $g_\ell(\eta_\alpha^\ell) - g_\ell(\theta_\alpha^0)$, step $(**)$ follows from $g_\ell(\theta_\alpha^\ell) - g_\ell(\theta_\alpha^0)$, $g_\ell(\eta_\alpha^\ell) - g_\ell(\theta_\alpha^0) \to 0$ due to (2.101), (2.102) and hence $\xi_\ell \to 0$, and step $(***)$ follows from (2.103) and (2.91).

Case 2: Calculate $\lim_{\ell\to\infty} I_i$ for $i \neq \alpha$.

For $i \neq \alpha$,

$$\begin{aligned}
&\frac{\exp\left(g_\ell(\theta_i^\ell)\right)}{\exp\left(g_\ell(\theta_\alpha^0)\right)} \\
=& \exp\left(\left\langle \theta_i^\ell - \theta_\alpha^0, \sqrt{N_\ell} z_\ell + N_\ell \nabla_\theta A(\theta_\alpha^0) \right\rangle - N_\ell \left(A(\theta_i^\ell) - A(\theta_\alpha^0) \right) \right) \\
=& \exp\left(-N_\ell \left(A(\theta_i^\ell) - A(\theta_\alpha^0) - \left\langle \theta_i^\ell - \theta_\alpha^0, \nabla_\theta A(\theta_\alpha^0) \right\rangle - \frac{1}{\sqrt{N_\ell}} \left\langle \theta_i^\ell - \theta_\alpha^0, z_\ell \right\rangle \right) \right) \\
\leq& \exp\left(-\frac{N_\ell}{2} \left(A(\theta_i^0) - A(\theta_\alpha^0) - \left\langle \theta_i^0 - \theta_\alpha^0, \nabla_\theta A(\theta_\alpha^0) \right\rangle \right) \right) \quad \text{for sufficiently large } \ell,
\end{aligned} \tag{2.105}$$

where the last inequality follows from $\lim_{\ell\to\infty} \frac{1}{\sqrt{N_\ell}} \left\langle \theta_i^\ell - \theta_\alpha^0, z_\ell \right\rangle = 0$ and

$$A(\theta_i^0) - A(\theta_\alpha^0) - \left\langle \theta_i^0 - \theta_\alpha^0, \nabla_\theta A(\theta_\alpha^0) \right\rangle > 0, \tag{2.106}$$

implied by strict convexity of $A(\theta)$ over Θ° due to full rank property of exponential family. Similarly, for sufficiently large ℓ,

$$\frac{\exp\left(g_\ell(\eta_i^\ell)\right)}{\exp\left(g_\ell(\theta_\alpha^0)\right)} \leq \exp\left(-\frac{N_\ell}{2}\left(A(\theta_i^0) - A(\theta_\alpha^0) - \left\langle \theta_i^0 - \theta_\alpha^0, \nabla_\theta A(\theta_\alpha^0)\right\rangle\right)\right). \tag{2.107}$$

It follows that for $i \neq \alpha$

$$\begin{aligned}
&\lim_{\ell\to\infty} |I_i| \\
\leq &\lim_{\ell\to\infty} p_i^\ell \left|\frac{\exp\left(g_\ell(\theta_i^\ell)\right) - \exp\left(g_\ell(\eta_i^\ell)\right)}{D_{N_\ell}(G_\ell, H_\ell)\exp\left(g_\ell(\theta_\alpha^0)\right)}\right| + \lim_{\ell\to\infty}\left|\frac{p_i^\ell - \pi_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)}\right| \frac{\exp\left(g_\ell(\eta_i^\ell)\right)}{\exp\left(g_\ell(\theta_\alpha^0)\right)} \\
\leq &p_i^0 \lim_{\ell\to\infty} \frac{\max\{\exp\left(g_\ell(\theta_i^\ell)\right), \exp\left(g_\ell(\eta_i^\ell)\right)\}}{\exp\left(g_\ell(\theta_\alpha^0)\right)} \left|\frac{g_\ell(\theta_i^\ell) - g_\ell(\eta_i^\ell)}{D_{N_\ell}(G_\ell, H_\ell)}\right| + |b_i| \lim_{\ell\to\infty} \frac{\exp\left(g_\ell(\eta_i^\ell)\right)}{\exp\left(g_\ell(\theta_\alpha^0)\right)} \\
\leq &\lim_{\ell\to\infty} \exp\left(-\frac{N_\ell}{2}\left(A(\theta_i^0) - A(\theta_\alpha^0) - \left\langle \theta_i^0 - \theta_\alpha^0, \nabla_\theta A(\theta_\alpha^0)\right\rangle\right)\right)\left(p_i^0\left|\frac{g_\ell(\theta_i^\ell) - g_\ell(\eta_i^\ell)}{D_{N_\ell}(G_\ell, H_\ell)}\right| + |b_i|\right),
\end{aligned} \tag{2.108}$$

where the second inequality follows by applying the mean value theorem on the first term and applying (2.91) to the second term, while the last inequality follows from (2.105) and (2.107).

Since

$$\begin{aligned}
&\limsup_{\ell\to\infty} \frac{1}{\sqrt{N_\ell}}\left|\frac{g_\ell(\theta_i^\ell) - g_\ell(\eta_i^\ell)}{D_{N_\ell}(G_\ell, H_\ell)}\right| \\
= &\limsup_{\ell\to\infty} \frac{1}{\sqrt{N_\ell}}\left|\frac{\left\langle \theta_i^\ell - \eta_i^\ell, \sqrt{N_\ell} z_\ell + N_\ell \nabla_\theta A(\theta_\alpha^0)\right\rangle - N_\ell\left(A(\theta_i^\ell) - A(\eta_i^\ell)\right)}{D_{N_\ell}(G_\ell, H_\ell)}\right| \\
\leq &\limsup_{\ell\to\infty} \frac{1}{\sqrt{N_\ell}}\left|\frac{\left\langle \sqrt{N_\ell}(\theta_i^\ell - \eta_i^\ell), z_\ell\right\rangle}{D_{N_\ell}(G_\ell, H_\ell)}\right| + \limsup_{\ell\to\infty}\left|\frac{-\sqrt{N_\ell}\left(A(\theta_i^\ell) - A(\theta_i^0) - \left\langle \theta_i^\ell - \eta_i^\ell, \nabla_\theta A(\theta_i^0)\right\rangle\right)}{D_{N_\ell}(G_\ell, H_\ell)}\right| \\
&\quad + \limsup_{\ell\to\infty}\left|\frac{\sqrt{N_\ell}\left\langle \theta_i^\ell - \eta_i^\ell, \nabla_\theta A(\theta_\alpha^0) - \nabla_\theta A(\theta_i^0)\right\rangle}{D_{N_\ell}(G_\ell, H_\ell)}\right| \\
= &\left|\left\langle a_i, \nabla_\theta A(\theta_\alpha^0) - \nabla_\theta A(\theta_i^0)\right\rangle\right|,
\end{aligned}$$

where the last step follows from (2.91) and (2.100). Then for sufficiently large ℓ

$$\left|\frac{g_\ell(\theta_i^\ell) - g_\ell(\eta_i^\ell)}{D_{N_\ell}(G_\ell, H_\ell)}\right| \leq \left(\left|\left\langle a_i, \nabla_\theta A(\theta_\alpha^0) - \nabla_\theta A(\theta_i^0)\right\rangle\right| + \frac{1}{\ell}\right)\sqrt{N_\ell}. \tag{2.109}$$

Plug (2.109) into (2.108), for any $i \neq \alpha$,

$$\begin{aligned}\lim_{\ell\to\infty}|I_i| \leq & \lim_{\ell\to\infty} e^{-\frac{N_\ell}{2}\left(A(\theta_i^0)-A(\theta_\alpha^0)-\left\langle\theta_i^0-\theta_\alpha^0,\nabla_\theta A(\theta_\alpha^0)\right\rangle\right)}\left(\left|\left\langle a_i,\nabla_\theta A(\theta_\alpha^0)-\nabla_\theta A(\theta_i^0)\right\rangle\right|+\frac{1}{\ell}\right)\sqrt{N_\ell}\\ =&0. \end{aligned} \tag{2.110}$$

Combine (2.96), (2.104) and (2.110), we see that (2.94) is established. This concludes the proof of the theorem. □

2.12.2 Proof of Theorem 2.5.14

Proof of Theorem 2.5.14: Step 1 (Proof by contradiction with subsequences)
This step is similar to the proof of Theorem 2.5.7. Suppose that (2.24) is not true. Then $\exists\{N_\ell\}_{\ell=1}^\infty$ subsequence of natural numbers tending to infinity such that

$$\lim_{r\to 0}\inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}}\frac{V(P_{G,N_\ell},P_{H,N_\ell})}{D_{N_\ell}(G,H)}\to 0 \quad \text{as } N_\ell\to\infty.$$

Then $\exists\{G_\ell\}_{\ell=1}^\infty,\{H_\ell\}_{\ell=1}^\infty\subset\mathcal{E}_{k_0}(\Theta)$ such that

$$\begin{cases} G_\ell\neq H_\ell & \forall\ell\\ D_{N_\ell}(G_\ell,G_0)\to 0, D_{N_\ell}(H_\ell,G_0)\to 0 & \text{as } \ell\to\infty\\ \frac{V(P_{G_\ell,N_\ell},P_{H_\ell,N_\ell})}{D_{N_\ell}(G_\ell,H_\ell)}\to 0 & \text{as } \ell\to\infty. \end{cases} \tag{2.111}$$

To see this, for each fixed ℓ, and thus fixed N_ℓ, $D_{N_\ell}(G,G_0)\to 0$ if and only if $W_1(G,G_0)\to 0$. Thus, there exist $G_\ell,H_\ell\in\mathcal{E}_{k_0}(\Theta)$ such that $G_\ell\neq H_\ell$, $D_{N_\ell}(G_\ell,G_0)\leq\frac{1}{\ell}$, $D_{N_\ell}(H_\ell,G_0)\leq\frac{1}{\ell}$ and

$$\frac{V(P_{G_\ell,N_\ell},P_{H_\ell,N_\ell})}{D_{N_\ell}(G_\ell,H_\ell)}\leq\lim_{r\to 0}\inf_{\substack{G,H\in B_{W_1}(G_0,r)\\ G\neq H}}\frac{V(P_{G,N_\ell},P_{H,N_\ell})}{D_{N_\ell}(G,H)}+\frac{1}{\ell},$$

thereby ensuring that (2.111) hold.

Write $G_0=\sum_{i=1}^{k_0}p_i^0\delta_{\theta_i^0}$. We may relabel the atoms of G_ℓ and H_ℓ such that $G_\ell=\sum_{i=1}^{k_0}p_i^\ell\delta_{\theta_i^\ell}$, $H_\ell=\sum_{i=1}^{k_0}\pi_i^\ell\delta_{\eta_i^\ell}$ with $\theta_i^\ell,\eta_i^\ell\to\theta_i^0$ and $p_i^\ell,\pi_i^\ell\to p_i^0$ for any $i\in[k_0]$. By subsequence argument if necessary, we may require $\{G_\ell\}_{\ell=1}^\infty$, $\{H_\ell\}_{\ell=1}^\infty$ additionally satisfy:

$$D_{N_\ell}(H_\ell,G_0)=\sum_{i=1}^{k_0}(\sqrt{N_\ell}\|\eta_i^\ell-\theta_i^0\|_2+|\pi_i^\ell-p_i^0|) \tag{2.112}$$

$$D_{N_\ell}(G_\ell, G_0) = \sum_{i=1}^{k_0} (\sqrt{N_\ell}\|\theta_i^\ell - \theta_i^0\|_2 + |p_i^\ell - p_i^0|) \tag{2.113}$$

and

$$\frac{\sqrt{N_\ell}\left(\theta_i^\ell - \eta_i^\ell\right)}{D_{N_\ell}(G_\ell, H_\ell)} \to a_i \in \mathbb{R}^s, \quad \frac{p_i^\ell - \pi_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \to b_i \in \mathbb{R}, \quad \forall 1 \le i \le k_0, \tag{2.114}$$

where the components of a_i are in $[-1, 1]$ and $\sum_{i=1}^{k_0} b_i = 0$. It also follows that at least one of a_i is not $\mathbf{0} \in \mathbb{R}^s$ or one of b_i is not 0. Let $\alpha \in \{1 \le i \le k_0 : a_i \ne \mathbf{0} \text{ or } b_i \ne 0\}$.

Step 2 (Transform the probability measure to support in $\mathbb{R}^s$)

Let $T_1 : (\mathfrak{X}, \mathcal{A}) \to (\mathbb{R}^s, \mathcal{B}(\mathbb{R}^s))$ be an arbitrary measurable map in this step. Extend T_1 to product space by $\bar{T}_1 : \mathfrak{X}^N \to \mathbb{R}^{Ns}$ by $\bar{T}_1\bar{x} = ((T_1x_1)^T, \ldots, (T_1x_N)^T)^T$ where any $\bar{x} \in \mathfrak{X}^N$ is partitioned into N blocks as $\bar{x} = (x_1, x_2, \ldots, x_N)$ with $x_i \in \mathfrak{X}$. Then one can easily verify that $(\bigotimes^N P_\theta) \circ \bar{T}_1^{-1} = \bigotimes^N (P_\theta \circ T_1^{-1})$, and hence for any $G \in \mathcal{E}_{k_0}(\Theta)$

$$P_{G,N} \circ \bar{T}_1^{-1} = \sum_{i=1}^{k_0} p_i (P_{\theta_i,N} \circ \bar{T}_1^{-1}) = \sum_{i=1}^{k_0} p_i \left(\bigotimes^N \left(P_{\theta_i} \circ T_1^{-1} \right) \right).$$

Further consider another measurable map $T_0 : (\mathbb{R}^{Ns}, \mathcal{B}(\mathbb{R}^{Ns})) \to (\mathbb{R}^s, \mathcal{B}(\mathbb{R}^s))$ defined by $T_0\bar{t} = \sum_{i=1}^N t_i$ where $\bar{t} \in \mathbb{R}^{Ns}$ is partitioned equally into N blocks $\bar{t} = (t_1^T, t_2^T, \ldots, t_N^T)^T \in \mathbb{R}^{Ns}$. Denote the induced probability measure on $\mathbb{R}^s$ under $T_0 \circ \bar{T}_1$ of the $P_{\theta,N}$ by $Q_{\theta,N} := \left(\bigotimes^N \left(P_\theta \circ T_1^{-1} \right) \right) \circ T_0^{-1}$. Then the induced probability measure under $T_0 \circ \bar{T}_1$ of the mixture $P_{G,N}$ is

$$P_{G,N} \circ \bar{T}_1^{-1} \circ T_0^{-1} = \sum_{i=1}^{k_0} p_i Q_{\theta_i,N} := Q_{G,N}.$$

Note the dependences of $\bar{T}_1$ and T_0 on N are both suppressed, so are the dependences on T_1 of $Q_{\theta,N}$ and $Q_{G,N}$.

Then by definition of total variation distance

$$V(P_{G,N}, P_{H,N}) \ge V(Q_{G,N}, Q_{H,N}), \quad \forall N, \forall\, T_1.$$

The above display and (2.111) yield

$$\lim_{\ell \to 0} \frac{V(Q_{G_\ell,N_\ell}, Q_{H_\ell,N_\ell})}{D_{N_\ell}(G_\ell, H_\ell)} = 0, \quad \forall T_1. \tag{2.115}$$

Step 3 (Application of the central limit theorem)

In the rest of proof specialize T_1 in step 2 to be T_α. Write $T = T_\alpha$ to simplify the notation in the rest of the proof. Let $\gamma > 0$ and $r \ge 1$ be the same as in Definition 2.5.13 of $T = T_\alpha$ with respect to

the finite set $\{\theta_i^0\}_{i=1}^{k_0}$ and define $\bar{\Theta}(G_0) := \bigcup_{i=1}^{k_0} B(\theta_i^0, \gamma)$. By subsequences if necessary, we may further require that G_ℓ satisfy $\theta_i^\ell \in B(\theta_i^0, \gamma)$ for all $i \in [k_0]$ and $N_\ell \geq r$.

Consider $\{X_i\}_{i=1}^{\infty} \overset{i.i.d.}{\sim} P_\theta$. Then $Y_\ell = \sum_{i=1}^{N_\ell} TX_i$ is distributed by probability measure Q_{θ,N_ℓ}, which has characteristic function $(\phi_T(\zeta|\theta))^{N_\ell}$. For $\theta \in \bar{\Theta}(G_0)$ by (A3) in the definition of admissible transform, by Fourier inversion theorem Q_{θ,N_ℓ} and Y_ℓ therefore have density $f_Y(y|\theta, N_\ell)$ w.r.t Lebesgue measure given by

$$f_Y(y|\theta, N_\ell) = \frac{1}{(2\pi)^s} \int_{\mathbb{R}^s} e^{-i\zeta^T y} (\phi_T(\zeta|\theta))^{N_\ell} d\zeta. \tag{2.116}$$

Then Q_{G_ℓ,N_ℓ} has density w.r.t. Lebesgue measure given by $\sum_{i=1}^{k_0} p_i f_Y(y|\theta_i^\ell, N_\ell)$, and thus

$$2V(Q_{G_\ell,N_\ell}, Q_{H_\ell,N_\ell}) = \int_{\mathbb{R}^s} \left| \sum_{i=1}^{k_0} p_i^\ell f_Y(y|\theta_i^\ell, N_\ell) - \sum_{i=1}^{k_0} \pi_i^\ell f_Y(y|\eta_i^\ell, N_\ell) \right| dy. \tag{2.117}$$

For Y_ℓ has density $f_Y(y|\theta, N_\ell)$, define $Z_\ell = (Y_\ell - N_\ell \lambda_\theta)/\sqrt{N_\ell}$. Note this transform from Y_ℓ to Z_ℓ depends on θ in the density of Y_ℓ. Then by the change of variable formula, Z_ℓ has density $f_Z(z|\theta, N_\ell)$ w.r.t. Lebesgue measure, given by

$$f_Z(z|\theta, N_\ell) = f_Y(\sqrt{N_\ell} z + N_\ell \lambda_\theta | \theta, N_\ell) N_\ell^{s/2},$$

or equivalently

$$f_Y(y|\theta, N_\ell) = f_Z((y - N_\ell \lambda_\theta)/\sqrt{N_\ell} | \theta, N_\ell)/N_\ell^{s/2}. \tag{2.118}$$

Now, applying the local central limit theorem (Lemma 2.16.1), $f_Z(z|\theta, N_\ell)$ converges uniformly in z to $f_{\mathcal{N}}(z|\theta)$ for every $\theta \in \bar{\Theta}(G_0)$. Next specialize to θ_α^0, and define

$$w_\ell = \sup \left\{ w \geq 0 : f_Z(z|\theta_\alpha^0, N_\ell) \geq \frac{1}{(2\pi)^{s/2}} \frac{1}{2\ell} \text{ for all } \|z\|_2 \leq w \right\}.$$

We use the convention that the supreme of $\emptyset$ is 0 in the above display. Because of the uniform convergence of $f_Z(z|\theta_\alpha^0, N_\ell)$ to $f_{\mathcal{N}}(z|\theta_\alpha^0)$, we have $w_\ell \to \infty$ when $\ell \to \infty$. It follows from (2.118) that $f_Y(y|\theta_\alpha^0, N_\ell) > 0$ on $B_\ell := \{y \in \mathbb{R}^s | y = \sqrt{N_\ell} z + N_\ell \lambda_{\theta_\alpha^0} \text{ for } \|z\|_2 \leq w_\ell\}$. Then by (2.117)

$$\begin{aligned}
&\frac{2V(Q_{G_\ell,N_\ell}, Q_{H_\ell,N_\ell})}{D_{N_\ell}(G_\ell, H_\ell)} \\
&= \int_{\mathbb{R}^s} \left| \sum_{i=1}^{k_0} \frac{p_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \left(f_Y(y|\theta_i^\ell, N_\ell) - f_Y(y|\eta_i^\ell, N_\ell) \right) + \sum_{i=1}^{k_0} \frac{p_i^\ell - \pi_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)} f_Y(y|\eta_i^\ell, N_\ell) \right| dy
\end{aligned}$$

$$
\begin{aligned}
&\geq \int_{B_\ell} \left| \sum_{i=1}^{k_0} \frac{p_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \frac{f_Y(y|\theta_i^\ell, N_\ell) - f_Y(y|\eta_i^\ell, N_\ell)}{f_Y(y|\theta_\alpha^0, N_\ell)} + \right. \\
&\qquad \left. \sum_{i=1}^{k_0} \frac{p_i^\ell - \pi_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \frac{f_Y(y|\eta_i^\ell, N_\ell)}{f_Y(y|\theta_\alpha^0, N_\ell)} \right| f_Y(y|\theta_\alpha^0, N_\ell) dy \\
&= \mathbb{E}_{\theta_\alpha^0} |F_\ell(Y_\ell)| \\
&= \mathbb{E}_{\theta_\alpha^0} |\Psi_\ell(Z_\ell)|, \qquad (2.119)
\end{aligned}
$$

where

$$
\begin{aligned}
&F_\ell(y) \\
&= \left(\sum_{i=1}^{k_0} \frac{p_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \frac{f_Y(y|\theta_i^\ell, N_\ell) - f_Y(y|\eta_i^\ell, N_\ell)}{f_Y(y|\theta_\alpha^0, N_\ell)} + \sum_{i=1}^{k_0} \frac{p_i^\ell - \pi_i^\ell}{D_{N_\ell}(G_\ell, H_\ell)} \frac{f_Y(y|\eta_i^\ell, N_\ell)}{f_Y(y|\theta_\alpha^0, N_\ell)} \right) \mathbf{1}_{B_\ell}(y),
\end{aligned}
$$

and

$$
\Psi_\ell(z) = F_\ell(\sqrt{N_\ell} z + N_\ell \lambda_{\theta_\alpha^0}).
$$

Observe if Z_ℓ has density $f_Z(z|\theta_\alpha^0, N_\ell)$, then Z_ℓ converges in distribution to $Z \sim \mathcal{N}(\mathbf{0}, \Lambda_{\theta_\alpha^0})$.

Step 4 (Application of a continuous mapping theorem)

Define $\Psi(z) = p_\alpha^0 \left(J_\lambda(\theta_\alpha^0) a_\alpha\right)^T \Lambda_{\theta_\alpha^0}^{-1} z + b_\alpha$, where $J_\lambda(\theta_\alpha^0) \in \mathbb{R}^{s \times q}$ is the Jacobian matrix of $\lambda(\theta)$ evaluated at θ_α^0. Suppose:

$$
\Psi_\ell(z_\ell) \to \Psi(z) \text{ for any sequence } z_\ell \to z \in \mathbb{R}^s, \qquad (2.120)
$$

a property to be verified later, then by Generalized Continuous Mapping Theorem ([WVdV96] Theorem 1.11.1), $\Psi_\ell(Z_\ell)$ convergence in distribution to $\Psi(Z)$. Apply Theorem 25.11 in [Bil96],

$$
\mathbb{E}|\Psi(Z)| \leq \liminf_{\ell \to \infty} \mathbb{E}_{\theta_\alpha^0} |\Psi_\ell(Z_\ell)| = 0,
$$

where the equality follows (2.119) and (2.115). Note that Λ_θ is positive definite (by (A1)) and $J_\lambda(\theta_\alpha^0)$ is of full column rank. In addition, by our choice of α, either a_α or b_α is non-zero. Hence, $\Psi(z)$ is a non-zero affine function of z. For such an $\Psi(z)$, $\mathbb{E}|\Psi(Z)|$ cannot be zero, which results in a contradiction. As a result, it remains in the proof to establish (2.120).

We will now impose the following technical claim and proceed to verify (2.120), while the proof of the claim will be given after the current proof.

Claim: For any $1 \leq i \leq k_0$, for any pair of sequences $\bar{\theta}_i^\ell, \bar{\eta}_i^\ell \in B(\theta_i^0, \gamma)$ and for any increasing

$\bar{N}_\ell \geq r$ satisfying $\sqrt{\bar{N}_\ell}\|\bar{\theta}_i^\ell - \theta_i^0\|_2, \sqrt{\bar{N}_\ell}\|\bar{\eta}_i^\ell - \theta_i^0\|_2 \to 0$ and $\bar{N}_\ell \to \infty$:

$$\begin{aligned} J(\bar{\theta}_i^\ell, \bar{\eta}_i^\ell, \bar{N}_\ell) :=& \bar{N}_\ell^{s/2} \sup_{y\in\mathbb{R}^s} \left| f_Y(y|\bar{\theta}_i^\ell, \bar{N}_\ell) - f_Y(y|\bar{\eta}_i^\ell, \bar{N}_\ell) - \sum_{j=1}^q \frac{\partial f_{\mathcal{N}}(y|\bar{\theta}_i^0, \bar{N}_\ell)}{\partial \theta^{(j)}} \left((\bar{\theta}_i^\ell)^{(j)} - (\bar{\eta}_i^\ell)^{(j)}\right) \right| \\ =& o(\sqrt{\bar{N}_\ell}\|\bar{\theta}_i^\ell - \bar{\eta}_i^\ell\|_2), \quad \text{as } \ell \to \infty, \end{aligned} \tag{2.121}$$

where $f_{\mathcal{N}}(y|\theta, N)$ is the density w.r.t. Lebesgue measure of $\mathcal{N}(N\lambda_\theta, N\Lambda_\theta)$ when Λ_θ is positive definite.

Step 5 (Verification of (2.120))
Write $D_\ell = D_{N_\ell}(G_\ell, H_\ell)$ for abbreviation in the remaining of this proof. Observe by the local central limit theorem (Lemma 2.16.1)

$$\left|f_Z(z_\ell|\theta_\alpha^0, N_\ell) - f_{\mathcal{N}}(z|\theta_\alpha^0)\right| \leq \sup_{z'\in\mathbb{R}^s} |f_Z(z'|\theta_\alpha^0, N_\ell) - f_{\mathcal{N}}(z'|\theta_\alpha^0)| + |f_{\mathcal{N}}(z_\ell|\theta_\alpha^0) - f_{\mathcal{N}}(z|\theta_\alpha^0)| \to 0,$$

as $\ell \to \infty$, which implies

$$\lim_{\ell\to\infty} f_Z(z_\ell|\theta_\alpha^0, N_\ell) = f_{\mathcal{N}}(z|\theta_\alpha^0). \tag{2.122}$$

Hereafter $\frac{\partial f_Y(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}|\theta_i^0, N_\ell)}{\partial\theta^{(j)}} := \frac{\partial f_Y(y|\theta_i^0, N_\ell)}{\partial\theta^{(j)}}\Big|_{y=\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}}$. Similar definition applies to $\frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}|\theta_i^0, N_\ell)}{\partial\theta^{(j)}}$. Then for each $i \in [k_0]$,

$$\begin{aligned} &\frac{1}{D_\ell}\frac{f_Y(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}|\theta_i^\ell, N_\ell) - f_Y(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}|\eta_i^\ell, N_\ell)}{f_Y(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}|\theta_\alpha^0, N_\ell)}\mathbf{1}_{B_\ell}(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}) \\ =&\frac{N_\ell^{s/2}}{D_\ell}\frac{f_Y(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}|\theta_i^\ell, N_\ell) - f_Y(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}|\eta_i^\ell, N_\ell)}{f_Z(z_\ell|\theta_\alpha^0, N_\ell)}\mathbf{1}_{E_\ell}(z_\ell) \\ \leq&\left(\frac{N_\ell^{s/2}}{D_\ell}\frac{\sum_{j=1}^q \frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell}z_\ell + N_\ell\lambda_{\theta_\alpha^0}|\theta_i^0, N_\ell)}{\partial\theta^{(j)}}\left((\theta_i^\ell)^{(j)} - (\eta_i^\ell)^{(j)}\right)}{f_Z(z_\ell|\theta_\alpha^0, N_\ell)} + \frac{J(\theta_i^\ell, \eta_i^\ell, N_\ell)}{D_\ell f_Z(z_\ell|\theta_\alpha^0, N_\ell)}\right)\mathbf{1}_{E_\ell}(z_\ell), \end{aligned} \tag{2.123}$$

where the first equality follows from (2.118) and where in the first equality $E_\ell = \{z \in \mathbb{R}^s | \|z\|_2 \leq w_\ell\}$. Observe that for any $i \in [k_0]$

$$\sqrt{N_\ell}\|\theta_i^\ell - \theta_i^0\| \to 0, \tag{2.124}$$

$$\sqrt{N_\ell}\|\eta_i^\ell - \theta_i^0\| \to 0 \tag{2.125}$$

by (2.112), (2.113) and (2.111). Then by applying (2.121) with $\bar{\theta}_i^\ell, \bar{\eta}_i^\ell, \bar{N}_\ell$ respectively be $\theta_i^\ell, \eta_i^\ell, N_\ell$,

and by (2.114),

$$\lim_{\ell\to\infty}\frac{J(\theta_i^\ell,\eta_i^\ell,N_\ell)}{D_\ell}\to 0,$$

which together with (2.122) yield

$$\lim_{\ell\to\infty}\frac{J(\theta_i^\ell,\eta_i^\ell,N_\ell)}{D_\ell f_Z(z_\ell|\theta_\alpha^0,N_\ell)}\mathbf{1}_{E_\ell}(z_\ell)\to 0. \tag{2.126}$$

Thus by (2.123) and (2.126)

$$\begin{aligned}
&\lim_{\ell\to\infty}\sum_{i=1}^{k_0}\frac{p_i^\ell}{D_\ell}\frac{f_Y(\sqrt{N_\ell}z_\ell+N_\ell\lambda_{\theta_\alpha^0}|\theta_i^\ell,N_\ell)-f_Y(\sqrt{N_\ell}z_\ell+N_\ell\lambda_{\theta_\alpha^0}|\theta_i^0,N_\ell)}{f_Y(\sqrt{N_\ell}z_\ell+N_\ell\lambda_{\theta_\alpha^0}|\theta_\alpha^0,N_\ell)}\mathbf{1}_{B_\ell}(\sqrt{N_\ell}z_\ell+N_\ell\lambda_{\theta_\alpha^0})\\
=&\sum_{i=1}^{k_0}p_i^0\lim_{\ell\to\infty}\left(\frac{N_\ell^{s/2}}{D_\ell}\frac{\sum_{j=1}^q\frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell}z_\ell+N_\ell\lambda_{\theta_\alpha^0}|\theta_i^0,N_\ell)}{\partial\theta^{(j)}}\left((\theta_i^\ell)^{(j)}-(\eta_i^\ell)^{(j)}\right)}{f_Z(z_\ell|\theta_\alpha^0,N_\ell)}\right)\mathbf{1}_{E_\ell}(z_\ell),
\end{aligned}\tag{2.127}$$

provided the right hand side exists.

Note that for each $j=1,\dots,q$, and any $\theta\in\bar{\Theta}(G_0)$, by a standard calculation for Gaussian density,

$$\begin{aligned}
&\frac{\partial f_{\mathcal{N}}(y|\theta,N)}{\partial\theta^{(j)}}\\
=&f_{\mathcal{N}}(y|\theta,N)\left(-\frac{1}{2}\det(\Lambda_\theta)^{-1}\frac{\partial\det(\Lambda_\theta)}{\partial\theta^{(j)}}+\right.\\
&\left.\left(\frac{\partial\lambda_\theta}{\partial\theta^{(j)}}\right)^T\Lambda_\theta^{-1}(y-N\lambda_\theta)-\frac{1}{2N}(y-N\lambda_\theta)^T\left(\frac{\partial\Lambda_\theta^{-1}}{\partial\theta^{(j)}}\right)(y-N\lambda_\theta)\right),
\end{aligned}$$

so we have

$$\begin{aligned}
&\frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell}z_\ell+N_\ell\lambda_{\theta_\alpha^0}|\theta_i^0,N_\ell)}{\partial\theta^{(j)}}\\
=&f_{\mathcal{N}}(\sqrt{N_\ell}z_\ell+N_\ell\lambda_{\theta_\alpha^0}|\theta_i^0,N_\ell)\left(-\frac{1}{2}\det\left(\Lambda_{\theta_i^0}\right)^{-1}\frac{\partial\det(\Lambda_{\theta_i^0})}{\partial\theta^{(j)}}+\right.\\
&\left(\frac{\partial\lambda_{\theta_i^0}}{\partial\theta^{(j)}}\right)^T\Lambda_{\theta_i^0}^{-1}(\sqrt{N_\ell}z_\ell+N_\ell(\lambda_{\theta_\alpha^0}-\lambda_{\theta_i^0}))-\\
&\left.\frac{1}{2}(z_\ell+\sqrt{N_\ell}(\lambda_{\theta_\alpha^0}-\lambda_{\theta_i^0}))^T\left(\frac{\partial\Lambda_{\theta_i^0}^{-1}}{\partial\theta^{(j)}}\right)(z_\ell+\sqrt{N_\ell}(\lambda_{\theta_\alpha^0}-\lambda_{\theta_i^0}))\right)\\
=&N_\ell^{-\frac{s}{2}}f_{\mathcal{N}}(z_\ell+\sqrt{N_\ell}(\lambda_{\theta_\alpha^0}-\lambda_{\theta_i^0})|\theta_i^0)\left(-\frac{1}{2}\frac{1}{\det(\Lambda_{\theta_i^0})}\frac{\partial\det(\Lambda_{\theta_i^0})}{\partial\theta^{(j)}}+\right.
\end{aligned}$$

$$\left(\frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}}\right)^T \Lambda_{\theta_i^0}^{-1}(\sqrt{N_\ell} z_\ell + N_\ell(\lambda_{\theta_\alpha^0} - \lambda_{\theta_i^0}))-$$
$$\frac{1}{2}(z_\ell + \sqrt{N_\ell}(\lambda_{\theta_\alpha^0} - \lambda_{\theta_i^0}))^T \left(\frac{\partial \Lambda_{\theta_i^0}^{-1}}{\partial \theta^{(j)}}\right)(z_\ell + \sqrt{N_\ell}(\lambda_{\theta_\alpha^0} - \lambda_{\theta_i^0}))\Bigg).$$

Thus, when $i \neq \alpha$,

$$\begin{aligned} & N_\ell^{\frac{s-1}{2}} \left| \frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell)}{\partial \theta^{(j)}} \right| \\ \leq & N_\ell^{-\frac{1}{2}} f_{\mathcal{N}}(z_\ell + \sqrt{N_\ell}(\lambda_{\theta_\alpha^0} - \lambda_{\theta_i^0}) | \theta_i^0) C(\theta_i^0, z) N_\ell \\ \longrightarrow & 0, \end{aligned} \tag{2.128}$$

where the inequality holds for sufficiently large ℓ, $C(\theta_i^0, z)$ is a constant that only depends on θ_i^0 and z, and the last step follows from $\lambda_{\theta_\alpha^0} \neq \lambda_{\theta_i^0}$ by condition 1) in the statement of theorem.

When $i = \alpha$,

$$\begin{aligned} & N_\ell^{\frac{s-1}{2}} \frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell)}{\partial \theta^{(j)}} \\ = & N_\ell^{-\frac{1}{2}} f_{\mathcal{N}}(z_\ell | \theta_\alpha^0) \left(-\frac{1}{2} \det\left(\Lambda_{\theta_i^0}\right)^{-1} \frac{\partial \det\left(\Lambda_{\theta_i^0}\right)}{\partial \theta^{(j)}} + \left(\frac{\partial \lambda_{\theta_\alpha^0}}{\partial \theta^{(j)}}\right)^T \Lambda_{\theta_\alpha^0}^{-1}(\sqrt{N_\ell} z_\ell) - \frac{1}{2} z_\ell^T \left(\frac{\partial \Lambda_{\theta_\alpha^0}^{-1}}{\partial \theta^{(j)}}\right) z_\ell \right) \\ \longrightarrow & f_{\mathcal{N}}(z | \theta_\alpha^0) \left(\frac{\partial \lambda_{\theta_\alpha^0}}{\partial \theta^{(j)}}\right)^T \Lambda_{\theta_\alpha^0}^{-1} z. \end{aligned} \tag{2.129}$$

Plug (2.128), and (2.129) into (2.127), and then combine with (2.122) and (2.114),

$$\begin{aligned} & \lim_{\ell \to \infty} \sum_{i=1}^{k_0} \frac{p_i^\ell}{D_\ell} \frac{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^\ell, N_\ell) - f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \eta_i^\ell, N_\ell)}{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_\alpha^0, N_\ell)} \mathbf{1}_{B_\ell}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0}) \\ = & p_\alpha^0 \sum_{j=1}^{q} a_\alpha^{(j)} \left(\frac{\partial \lambda_{\theta_\alpha^0}}{\partial \theta^{(j)}}\right)^T \Lambda_{\theta_\alpha^0}^{-1} z \\ = & p_\alpha^0 \left(J_\lambda(\theta_\alpha^0) a_\alpha\right)^T \Lambda_{\theta_\alpha^0}^{-1} z. \end{aligned} \tag{2.130}$$

Next, we turn to the second summation in the definition of Ψ_ℓ in a similar fashion. By (2.118),

$$\begin{aligned} & \frac{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \eta_i^\ell, N_\ell)}{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_\alpha^0, N_\ell)} \mathbf{1}_{B_\ell}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0}) \\ = & N_\ell^{s/2} \frac{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \eta_i^\ell, N_\ell)}{f_Z(z_\ell | \theta_\alpha^0, N_\ell)} \mathbf{1}_{E_\ell}(z_\ell) \end{aligned}$$

$$\leq N_\ell^{s/2} \frac{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell) + \sum_{j=1}^q \frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell)}{\partial \theta^{(j)}} \left((\theta_i^\ell)^{(j)} - (\theta_i^0)^{(j)} \right)}{f_Z(z_\ell | \theta_\alpha^0, N_\ell)} \mathbf{1}_{E_\ell}(z_\ell)$$
$$+ \frac{J(\eta_i^\ell, \theta_i^0, N_\ell)}{f_Z(z_\ell | \theta_\alpha^0, N_\ell)} \mathbf{1}_{E_\ell}(z_\ell). \tag{2.131}$$

Due to (2.125), by applying (2.121) with $\bar{\theta}_i^\ell, \bar{\eta}_i^\ell, \bar{N}_\ell$ respectively be $\eta_i^\ell, \theta_i^0, N_\ell$, and by (2.111),

$$\lim_{\ell \to \infty} J(\eta_i^\ell, \theta_i^0, N_\ell) \to 0,$$

which together with (2.122) yield

$$\lim_{\ell \to \infty} \frac{J(\eta_i^\ell, \theta_i^0, N_\ell)}{f_Z(z_\ell | \theta_\alpha^0, N_\ell)} \mathbf{1}_{E_\ell}(z_\ell) \to 0. \tag{2.132}$$

Moreover for any $i \in [k_0]$,

$$N_\ell^{s/2} \sum_{j=1}^q \frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell)}{\partial \theta^{(j)}} \left((\theta_i^\ell)^{(j)} - (\theta_i^0)^{(j)} \right)$$
$$\leq \max_{1 \leq j \leq q} N_\ell^{(s-1)/2} \left| \frac{\partial f_{\mathcal{N}}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell)}{\partial \theta^{(j)}} \right| \sqrt{q} \sqrt{N_\ell} \| \theta_i^\ell - \theta_i^0 \|_2$$
$$\to 0. \tag{2.133}$$

by (2.128) and (2.129) and (2.124).

Combining (2.131), (2.132), (2.133) and (2.114),

$$\lim_{\ell \to \infty} \sum_{i=1}^{k_0} \frac{p_i^\ell - \pi_i^\ell}{D_\ell} \frac{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell)}{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_\alpha^0, N_\ell)} \mathbf{1}_{B_\ell}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0})$$
$$= \sum_{i=1}^{k_0} b_i \lim_{\ell \to \infty} N_\ell^{s/2} \frac{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell)}{f_Z(z_\ell | \theta_\alpha^0, N_\ell)} \mathbf{1}_{E_\ell}(z_\ell)$$
$$= \sum_{i=1}^{k_0} b_i \lim_{\ell \to \infty} \frac{f_Z(z_\ell + \sqrt{N_\ell}(\lambda_{\theta_\alpha^0} - \lambda_{\theta_i^0}) | \theta_i^0, N_\ell)}{f_Z(z_\ell | \theta_\alpha^0, N_\ell)} \mathbf{1}_{E_\ell}(z_\ell) \tag{2.134}$$

where the last step is due to (2.118).

When $i = \alpha$, the term in the preceding display equals to $\mathbf{1}_{E_\ell}(z_\ell)$, which converges to 1 as $\ell \to \infty$. When $i \neq \alpha$,

$$|f_Z(\sqrt{N_\ell}(\lambda_{\theta_\alpha^0} - \lambda_{\theta_i^0}) + z_\ell | \theta_i^0, N_\ell)|$$
$$\leq \sup_{z' \in \mathbb{R}^s} |f_Z(z' | \theta_i^0, N_\ell) - f_{\mathcal{N}}(z' | \theta_i^0)| + f_{\mathcal{N}}(\sqrt{N_\ell}(\lambda_{\theta_\alpha^0} - \lambda_{\theta_i^0}) + z_\ell | \theta_i^0)$$

$$\to 0, \tag{2.135}$$

where the last step follows from Lemma 2.16.1 and $\lambda_{\theta_\alpha^0} \neq \lambda_{\theta_i^0}$ by condition 1) in the statement of the theorem. Plug (2.135) and (2.122) into (2.134):

$$\lim_{\ell\to\infty} \sum_{i=1}^{k_0} \frac{p_i^\ell - \pi_i^\ell}{D_\ell} \frac{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_i^0, N_\ell)}{f_Y(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0} | \theta_\alpha^0, N_\ell)} \mathbf{1}_{B_\ell}(\sqrt{N_\ell} z_\ell + N_\ell \lambda_{\theta_\alpha^0}) = b_\alpha. \tag{2.136}$$

Finally, combining (2.136) and (2.130) to obtain

$$\lim_{\ell\to\infty} \Psi_\ell(z_\ell) = \Psi(z) = p_\alpha^0 \left(J_\lambda(\theta_\alpha^0) a_\alpha\right)^T \Lambda_{\theta_\alpha^0}^{-1} z + b_\alpha.$$

Thus, (2.120) is established, so we can conclude the proof of the theorem. □

Proof of Claim (2.121)**:** We will write $\theta_i^\ell, \eta_i^\ell, N_\ell$ respectively for $\bar{\theta}_i^\ell, \bar{\eta}_i^\ell, \bar{N}_\ell$ in this proof. But $\theta_i^\ell, \eta_i^\ell, N_\ell$ in this proof are generic variables and might not necessarily be the same as in the proof of Theorem 2.5.14.

For any $\theta \in \bar{\Theta}(G_0)$, by condition (A1) $\nabla_\zeta \, \phi(\zeta|\theta)|_{\zeta=\mathbf{0}} = \boldsymbol{i}\lambda_\theta$, and $\mathbf{Hess}_\zeta \, \phi(\zeta|\theta)|_{\zeta=\mathbf{0}} = \boldsymbol{i}^2 \left(\Lambda_\theta + \lambda_\theta \lambda_\theta^T\right)$ exist, and by condition (A2) $\frac{\partial \lambda_\theta}{\partial \theta^{(j)}}$ and $\frac{\partial \Lambda_\theta}{\partial \theta^{(j)}}$ exist. Then, with condition (A1) it follows from Pratt's Lemma that $\frac{\partial f_{\mathcal{N}}(y|\theta,N)}{\partial \theta^{(j)}}$ exists and is given by

$$\frac{\partial f_{\mathcal{N}}(y|\theta,N)}{\partial \theta^{(j)}} = \frac{1}{(2\pi)^s} \int_{\mathbb{R}^s} e^{-\boldsymbol{i}\zeta^T y} \exp\left(\boldsymbol{i} N \zeta^T \lambda_\theta - \frac{N}{2} \zeta^T \Lambda_\theta \zeta\right) \left(\boldsymbol{i} N \zeta^T \frac{\partial \lambda_\theta}{\partial \theta^{(j)}} - \frac{N}{2} \zeta^T \frac{\partial \Lambda_\theta}{\partial \theta^{(j)}} \zeta\right) d\zeta. \tag{2.137}$$

Plugging the Fourier inversion formula (2.116) and (2.137) into (2.121), and noting $|e^{-\boldsymbol{i}\zeta^T y}| \leq 1$ for all $y \in \mathbb{R}^s$, for sufficiently large ℓ we obtain

$$\begin{aligned} J(\theta_i^\ell, \eta_i^\ell, N_\ell) \leq & \frac{N_\ell^{s/2}}{(2\pi)^s} \int_{\mathbb{R}^s} \Big| (\phi_T(\zeta|\theta_i^\ell))^{N_\ell} - (\phi_T(\zeta|\eta_i^\ell))^{N_\ell} - \\ & -N_\ell e^{\boldsymbol{i} N_\ell \zeta^T \lambda_{\theta_i^0} - \frac{N_\ell}{2} \zeta^T \Lambda_{\theta_i^0} \zeta} \sum_{j=1}^{q} \left((\theta_i^\ell)^{(j)} - (\eta_i^\ell)^{(j)}\right) \left(\boldsymbol{i}\zeta^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}} - \frac{1}{2} \zeta^T \frac{\partial \Lambda_{\theta_i^0}}{\partial \theta^{(j)}} \zeta\right) \Big| \, d\zeta \\ \leq & \check{J}_\ell + \hat{J}_\ell, \end{aligned}$$

where

$$\begin{aligned} \check{J}_\ell := & \frac{N_\ell^{s/2}}{(2\pi)^s} \int_{\mathbb{R}^s} \Big| (\phi_T(\zeta|\theta_i^\ell))^{N_\ell} - (\phi_T(\zeta|\eta_i^\ell))^{N_\ell} - \\ & N_\ell \left(\phi_T(\zeta|\theta_i^0)\right)^{N_\ell - 1} \sum_{j=1}^{q} \left((\theta_i^\ell)^{(j)} - (\eta_i^\ell)^{(j)}\right) \frac{\partial \phi_T(\zeta|\theta_i^0)}{\partial \theta^{(j)}} \Big| \, d\zeta, \end{aligned}$$

and

$$\hat{J}_\ell := N_\ell^{s/2+1}\frac{1}{(2\pi)^s}\sum_{j=1}^{q}\left|(\theta_i^\ell)^{(j)}-(\eta_i^\ell)^{(j)}\right|\int_{\mathbb{R}^s}\left|\left(\phi_T(\zeta|\theta_i^0)\right)^{N_\ell-1}\frac{\partial\phi_T(\zeta|\theta_i^0)}{\partial\theta^{(j)}}-\right.$$
$$\left.-\exp\left(\boldsymbol{i}N_\ell\zeta^T\lambda_{\theta_i^0}-\frac{N_\ell}{2}\zeta^T\Lambda_{\theta_i^0}\zeta\right)\left(\boldsymbol{i}\zeta^T\frac{\partial\lambda_{\theta_i^0}}{\partial\theta^{(j)}}-\frac{1}{2}\zeta^T\frac{\partial\Lambda_{\theta_i^0}}{\partial\theta^{(j)}}\zeta\right)\right|d\zeta$$

We will show in the sequel that $\check{J}_\ell = o(\sqrt{N_\ell}\|\theta_i^\ell-\eta_i^\ell\|_2)$ in Step 1 and show $\hat{J}_\ell = o(\sqrt{N_\ell}\|\theta_i^\ell-\eta_i^\ell\|_2)$ in Step 2, thereby establishing (2.121).

Step 1 (Prove $\check{J}_\ell = o(\sqrt{N_\ell}\|\theta_i^\ell-\eta_i^\ell\|_2)$)

By Condition (A3) and Lemma 2.10.3 b),

$$\check{J}_\ell \le \frac{N_\ell^{s/2}}{(2\pi)^s}\int_{\mathbb{R}^s}\left|q\sum_{1\le j,\beta\le q}\left(\|\theta_i^\ell-\theta_i^0\|_2+\|\eta_i^\ell-\theta_i^0\|_2\right)\|\theta_i^\ell-\eta_i^\ell\|_2 R_1(\zeta;\theta_i^0,\theta_i^\ell,\eta_i^\ell,j,\beta)\right|d\zeta, \tag{2.138}$$

where with $\theta_\ell(t_1,t_2)=\theta_i^0+t_2(\eta_i^\ell+t_1(\theta_i^\ell-\eta_i^\ell))$

$$\begin{aligned}
&R_1(\zeta;\theta_i^0,\theta_i^\ell,\eta_i^\ell,j,\beta),\\
=&\int_0^1\int_0^1\left|N_\ell(N_\ell-1)\left(\phi_T(\zeta|\theta_\ell(t_1,t_2))\right)^{N_\ell-2}\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(\beta)}}+\right.\\
&\left.+N_\ell\left(\phi_T(\zeta|\theta_\ell(t_1,t_2))\right)^{N_\ell-1}\frac{\partial^2\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}\partial\theta^{(\beta)}}\right|dt_2dt_1.
\end{aligned}$$

Then

$$\begin{aligned}
&\int_{\mathbb{R}^s}\left|R_1(\zeta;\theta_i^0,\theta_i^\ell,\eta_i^\ell,j,\beta)\right|d\zeta\\
\le&N_\ell\int_{\mathbb{R}^s}\int_0^1\int_0^1\left|\phi_T(\zeta|\theta_\ell(t_1,t_2))\right|^{N_\ell-2}\times\\
&\left(N_\ell\left|\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(\beta)}}\right|+\left|\frac{\partial^2\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}\partial\theta^{(\beta)}}\right|\right)dt_2dt_1d\zeta\\
=&N_\ell\int_0^1\int_0^1\int_{\mathbb{R}^s}\left|\phi_T(\zeta|\theta_\ell(t_1,t_2))\right|^{N_\ell-2}\times\\
&\left(N_\ell\left|\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(\beta)}}\right|+\left|\frac{\partial^2\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}\partial\theta^{(\beta)}}\right|\right)d\zeta dt_2dt_1\\
=:&N_\ell R_2(\theta_i^0,\theta_i^\ell,\eta_i^\ell,j,\beta),
\end{aligned} \tag{2.139}$$

where the first inequality follows from the fact that $|\phi_T(\zeta|\theta_\ell(t_1,t_2)))|\le 1$, and the last inequality fol-

lows from Condition (A3), Tonelli Theorem and the joint Lebesgue measurability of $\phi_T(\zeta|\theta_\ell(t_1,t_2))$, $\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}$ and $\left|\frac{\partial^2\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}\partial\theta^{(\beta)}}\right|$, as functions of ζ, t_1 and t_2 by Lemma 4.51 of [AK06].

Then following (2.138) and (2.139),

$$
\begin{aligned}
&\check{J}_\ell \\
\leq& C(q,s)N_\ell^{s/2+1}\left\|\theta_i^\ell-\eta_i^\ell\right\|_2\left(\left\|\theta_i^\ell-\theta_i^0\right\|_2+\left\|\eta_i^\ell-\theta_i^0\right\|_2\right)\max_{1\leq j,\beta\leq q}R_2(\theta_i^0,\theta_i^\ell,\eta_i^\ell,j,\beta)\\
=& C(q,s)\sqrt{N_\ell}\left\|\theta_i^\ell-\eta_i^\ell\right\|_2\sqrt{N_\ell}\left(\left\|\theta_i^\ell-\theta_i^0\right\|_2+\left\|\eta_i^\ell-\theta_i^0\right\|_2\right)\times\\
&\max_{1\leq j,\beta\leq q}\int_0^1\int_0^1\int_{\mathbb{R}^s}\left|\phi_T\left(\frac{\bar\zeta}{\sqrt{N_\ell}}\middle|\theta_\ell(t_1,t_2)\right)\right|^{N_\ell-2}\left(N_\ell\left|\frac{\partial\phi_T(\frac{\bar\zeta}{\sqrt{N_\ell}}|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}\frac{\partial\phi_T(\frac{\bar\zeta}{\sqrt{N_\ell}}|\theta_\ell(t_1,t_2))}{\partial\theta^{(\beta)}}\right|\right.\\
&\left.+\left|\frac{\partial^2\phi_T(\frac{\bar\zeta}{\sqrt{N_\ell}}|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}\partial\theta^{(\beta)}}\right|\right)d\bar\zeta dt_2dt_1, \qquad (2.140)
\end{aligned}
$$

where in the first inequality $C(q,s)$ is some constant that depends on q and s, and where the second equality follows from (2.139) and changing variable with $\bar\zeta=\sqrt{N_\ell}\zeta$. Denote the integrand in the last display by $E_{j,\beta}(\bar\zeta,t_1,t_2)$.

Observe $f_Y(y|\theta_\ell(t_1,t_2),r)$ exists and has upper bound $\frac{1}{(2\pi)^s}\int_{\mathbb{R}^s}|\phi_T(\zeta|\theta_\ell(t_1,t_2))|^r d\zeta\leq C(s)U_2(\theta_i^0)$ by condition (A3). Then invoking Lemma 2.16.2, for $\|\zeta\|_2\leq 1$,

$$
|\phi_T(\zeta|\theta_\ell(t_1,t_2))|^r\leq\exp\left(-\frac{C(s)\|\zeta\|_2^2}{(\lambda_{\max}(\Lambda_{\theta_t(\ell)})+1)\lambda_{\max}^{s-1}(\Lambda_{\theta_t(\ell)}))U_2^2(\theta_i^0)}\right)\leq\exp\left(-\frac{C(s)\|\zeta\|_2^2}{U_3(\theta_i^0)U_2^2(\theta_i^0)}\right), \qquad (2.141)
$$

where the last step follows from $(\lambda_{\max}(\Lambda_{\theta_t(\ell)})+1)\lambda_{\max}^{s-1}(\Lambda_{\theta_t(\ell)})\leq U_3(\theta_i^0)$ by condition (A1) with $U_3(\theta_i^0)$ being some constant that depends on θ_i^0.

Moreover, by the mean value theorem and condition (A3): $\forall\|\zeta\|_2<1$

$$
\begin{aligned}
\left|\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}\right|&=\left|\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}-\frac{\partial\phi_T(0|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}\right|\\
&\leq\|\zeta\|_2\sup_{\|\zeta\|_2<1}\left\|\nabla_\zeta\frac{\partial\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}}\right\|_2\leq\sqrt{s}U_1(\theta_i^0)\|\zeta\|_2. \qquad (2.142)
\end{aligned}
$$

Then

$$
\begin{aligned}
&\int_{\|\bar\zeta\|_2<\sqrt{N_\ell}}E_{j,\beta}(\bar\zeta,t_1,t_2)d\bar\zeta\\
\leq&\int_{\|\bar\zeta\|_2<\sqrt{N_\ell}}\exp\left(-\frac{C(s)\|\bar\zeta\|_2^2}{rU_3(\theta_i^0)U_2^2(\theta_i^0)}\frac{N_\ell-2}{N_\ell}\right)\left(\left(\sqrt{s}U_1(\theta_i^0)\right)^2\|\bar\zeta\|_2^2+U_2(\theta_i^0)\right)d\bar\zeta\\
\leq&\int_{\mathbb{R}^s}\exp\left(-\frac{C(s)\|\bar\zeta\|_2^2}{2rU_3(\theta_i^0)U_2^2(\theta_i^0)}\right)\left(\left(\sqrt{s}U_1(\theta_i^0)\right)^2\|\bar\zeta\|_2^2+U_2(\theta_i^0)\right)d\bar\zeta
\end{aligned}
$$

$$=C(s,r,\theta_i^0), \tag{2.143}$$

where the first inequality follows from (2.141) and (2.142).

Let $\eta := \sup_{\|\zeta\|_2\geq 1}|\phi_T(\zeta|\theta_i^0)|$. Since the density $f_Y(y|\theta_i^0,r)$ w.r.t. Lebesgue exists and has characteristic function $\phi_T^r(\zeta|\theta_i^0)$, $\phi_T^r(\zeta|\theta_i^0)\to 0$ as $\|\zeta\|_2\to\infty$ by Riemann–Lebesgue lemma. It follows η is actually a maximum. Moreover, the existence of the density $f_Y(y|\theta_i^0,r)$ w.r.t. Lebesgue, together with Lemma 4 in Section 1, Chapter XV of [Fel08], yield $|\phi_T(\zeta|\theta_i^0)|^r<1$ when $\zeta\neq\mathbf{0}$. It follows that $\eta<1$. By mean value theorem and (A3)

$$\sup_{\zeta\in\mathbb{R}^s}|\phi_T(\zeta|\theta_\ell(t_1,t_2))-\phi_T(\zeta|\theta_i^0)|\leq\sqrt{q}U_1(\theta_i^0)\|\theta_i^\ell-\theta_i^0\|_2,$$

which further implies $\sup_{t\in[0,1]}\sup_{\|\zeta\|_2\geq 1}|\phi_T(\zeta|\theta_\ell(t_1,t_2))|<\eta+\frac{1-\eta}{2}:=\eta'<1$ for sufficiently large ℓ.

Then for sufficiently large ℓ,

$$\begin{aligned}
&\int_{\|\bar\zeta\|_2\geq\sqrt{N_\ell}}E_{j,\beta}(\bar\zeta,t_1,t_2)d\bar\zeta\\
\leq&(\eta')^{N_\ell-2-r}\int_{\mathbb{R}^s}\left|\phi_T\left(\frac{\bar\zeta}{\sqrt{N_\ell}}\middle|\theta_\ell(t_1,t_2)\right)\right|^r\left(N_\ell U_1^2(\theta_i^0)+\left|\frac{\partial^2\phi_T(\frac{\bar\zeta}{\sqrt{N_\ell}}|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}\partial\theta^{(\beta)}}\right|\right)d\bar\zeta\\
\leq&(\eta')^{N_\ell-2-r}N_\ell^{s/2}(N_\ell U_1^2(\theta_i^0)+1)\int_{\mathbb{R}^s}|\phi_T(\zeta|\,\theta_\ell(t_1,t_2))|^r\left(1+\left|\frac{\partial^2\phi_T(\zeta|\theta_\ell(t_1,t_2))}{\partial\theta^{(j)}\partial\theta^{(\beta)}}\right|\right)d\zeta\\
\leq&(\eta')^{N_\ell-2-r}N_\ell^{s/2}\left(N_\ell U_1^2(\theta_i^0)+1\right)U_2(\theta_i^0),
\end{aligned}\tag{2.144}$$

where the first inequality follows from the definition of η' and condition (A3), and the last inequality follows from condition (A3). (2.143) and (2.144) immediately imply for any j, β:

$$\limsup_{\ell\to\infty}\int_0^1\int_0^1\int_{\mathbb{R}^s}E_{j,\beta}(\bar\zeta,t_1,t_2)d\bar\zeta dt_2dt_1<\infty. \tag{2.145}$$

The above display together with (2.140) and the conditions $\sqrt{N_\ell}\|\theta_i^\ell-\theta_i^0\|_2,\sqrt{N_\ell}\|\eta_i^\ell-\theta_i^0\|_2\to 0$ yield $\check J_\ell=o(\sqrt{N_\ell}\|\theta_i^\ell-\eta_i^\ell\|_2)$.

Step 2 (Prove $\hat J_\ell=o(\sqrt{N_\ell}\|\theta_i^\ell-\eta_i^\ell\|_2)$). A large portion of the proof borrows ideas from Theorem 2 in Chapter XV, Section 5 of [Fel08].

Observe

$$\hat J_\ell\leq\sqrt{N_\ell}\|\theta_i^\ell-\eta_i^\ell\|_2\frac{\sqrt{q}}{(2\pi)^s}\max_{1\leq j\leq q}K_\ell(j) \tag{2.146}$$

where as before by a change of variable, $\bar{\zeta} = \sqrt{N_\ell}\zeta$,

$$\begin{aligned}
K_\ell(j) :=& N_\ell^{\frac{s+1}{2}} \int_{\mathbb{R}^s} \left| \left(\phi_T(\zeta|\theta_i^0)\right)^{N_\ell - 1} \frac{\partial \phi_T(\zeta|\theta_i^0)}{\partial \theta^{(j)}} - \right. \\
& \left. - \exp\left(\boldsymbol{i} N_\ell \zeta^T \lambda_{\theta_i^0} - \frac{N_\ell}{2} \zeta^T \Lambda_{\theta_i^0} \zeta \right) \left(\boldsymbol{i} \zeta^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}} - \frac{1}{2} \zeta^T \frac{\partial \Lambda_{\theta_i^0}}{\partial \theta^{(j)}} \zeta \right) \right| d\zeta \\
=& N_\ell^{\frac{s+1}{2}} \int_{\mathbb{R}^s} \left| \left(e^{-\boldsymbol{i}\zeta^T \lambda_{\theta_i^0}} \phi_T(\zeta|\theta_i^0) \right)^{N_\ell - 1} \frac{\partial \phi_T(\zeta|\theta_i^0)}{\partial \theta^{(j)}} - \right. \\
& \left. - \exp\left(\boldsymbol{i} \zeta^T \lambda_{\theta_i^0} - \frac{N_\ell}{2} \zeta^T \Lambda_{\theta_i^0} \zeta \right) \left(\boldsymbol{i} \zeta^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}} - \frac{1}{2} \zeta^T \frac{\partial \Lambda_{\theta_i^0}}{\partial \theta^{(j)}} \zeta \right) \right| d\zeta \\
=& \int_{\mathbb{R}^s} \sqrt{N_\ell} \left| \left(e^{-\frac{\boldsymbol{i}}{\sqrt{N_\ell}} \bar{\zeta}^T \lambda_{\theta_i^0}} \phi_T\left(\frac{\bar{\zeta}}{\sqrt{N_\ell}} | \theta_i^0 \right) \right)^{N_\ell - 1} \frac{\partial \phi_T(\frac{\bar{\zeta}}{\sqrt{N_\ell}}|\theta_i^0)}{\partial \theta^{(j)}} - \right. \\
& \left. - \exp\left(\frac{\boldsymbol{i}}{\sqrt{N_\ell}} \bar{\zeta}^T \lambda_{\theta_i^0} - \frac{1}{2} \bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta} \right) \left(\frac{\boldsymbol{i}}{\sqrt{N_\ell}} \bar{\zeta}^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}} - \frac{1}{2N_\ell} \bar{\zeta}^T \frac{\partial \Lambda_{\theta_i^0}}{\partial \theta^{(j)}} \bar{\zeta} \right) \right| d\bar{\zeta}. \quad (2.147)
\end{aligned}$$

Denote the integrand in the above display by A. Since $\lambda_{\theta_i^0}$ and $\Lambda_{\theta_i^0}$ exist, $e^{-\boldsymbol{i}\zeta^T \lambda_{\theta_i^0}} \phi_T(\zeta|\theta_i^0)$ is twice continuously differentiable on $\mathbb{R}^s$, with gradient being $\mathbf{0}$ and Hessian being $\boldsymbol{i}^2 \Lambda_{\theta_i^0}$ at $\zeta = 0$. Then by Taylor Theorem,

$$\left| e^{-i\zeta^T \lambda_{\theta_i^0}} \phi_T(\zeta|\theta_i^0) \right| < \exp\left(-\frac{1}{4} \zeta^T \Lambda_{\theta_i^0} \zeta \right) \quad \text{if } 0 < \|\zeta\|_2 < \gamma_1, \quad (2.148)$$

for sufficient small $0 < \gamma_1 < 1$ and

$$\left(e^{-\frac{\boldsymbol{i}}{\sqrt{N_\ell}} \bar{\zeta}^T \lambda_{\theta_i^0}} \phi_T\left(\frac{\bar{\zeta}}{\sqrt{N_\ell}} | \theta_i^0 \right) \right)^{N_\ell - 1} \to \exp\left(-\frac{1}{2} \bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta} \right). \quad (2.149)$$

Let $\eta'' := \sup_{\|\zeta\|_2 \geq \gamma_1} |\phi(\zeta|\theta_0)|$. By the same reasoning of $\eta < 1$ in Step 1, $\eta'' < 1$. Then for any $a > 0$,

$$\int_{\mathbb{R}^s} A d\bar{\zeta} = \int_{\|\bar{\zeta}\|_2 \leq a} A d\bar{\zeta} + \int_{a < \|\bar{\zeta}\|_2 < \gamma_1 \sqrt{N_\ell}} A d\bar{\zeta} + \int_{\|\bar{\zeta}\|_2 \geq \gamma_1 \sqrt{N_\ell}} A d\bar{\zeta}. \quad (2.150)$$

Then, as $\ell \to \infty$

$$\begin{aligned}
& \int_{\|\bar{\zeta}\|_2 \geq \gamma_1 \sqrt{N_\ell}} A d\bar{\zeta} \\
\leq & (\eta'')^{N_\ell - 1 - r} \sqrt{N_\ell} \int_{\mathbb{R}^s} \left| \phi_T\left(\frac{\bar{\zeta}}{\sqrt{N_\ell}} \middle| \theta_i^0 \right) \right|^r U_1(\theta_i^0) d\bar{\zeta} \\
& + \sqrt{N_\ell} \int_{\|\bar{\zeta}\|_2 \geq \gamma_1 \sqrt{N_\ell}} \exp\left(-\frac{1}{2} \bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta} \right) \left(\frac{1}{\sqrt{N_\ell}} \left| \bar{\zeta}^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}} \right| + \frac{1}{2N_\ell} \left| \bar{\zeta}^T \frac{\partial \Lambda_{\theta_i^0}}{\partial \theta^{(j)}} \bar{\zeta} \right| \right) d\bar{\zeta}
\end{aligned}$$

$$
\begin{aligned}
&= (\eta'')^{N_\ell - 1 - r} N_\ell^{\frac{s+1}{2}} U_1(\theta_i^0) \int_{\mathbb{R}^s} \left|\phi_T\left(\zeta|\,\theta_i^0\right)\right|^r d\zeta \\
&\quad + \int_{\|\bar{\zeta}\|_2 \geq \gamma_1 \sqrt{N_\ell}} \exp\left(-\frac{1}{2}\bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta}\right) \left(\left|\bar{\zeta}^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}}\right| + \frac{1}{2\sqrt{N_\ell}} \left|\bar{\zeta}^T \frac{\partial \Lambda_{\theta_i^0}}{\partial \theta^{(j)}} \bar{\zeta}\right|\right) d\bar{\zeta} \\
&\to 0,
\end{aligned} \tag{2.151}
$$

where the first inequality follows from condition (A3) and the definition of η'', and the last step follows from $\eta'' < 1$ and condition (A3).

By condition (A2), $\frac{\partial \phi_T(\zeta|\theta_i^0)}{\partial \theta^{(j)}}$ as a function of ζ has gradient at 0: $\boldsymbol{i} \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}}$. Then by Taylor Theorem:

$$
\sqrt{N_\ell} \frac{\partial \phi_T(\frac{\bar{\zeta}}{\sqrt{N_\ell}}|\theta_i^0)}{\partial \theta^{(j)}} \to \boldsymbol{i} \bar{\zeta}^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}}. \tag{2.152}
$$

Moreover, specialize $t = 0$ in (2.142): $\forall \|\zeta\|_2 < 1$

$$
\left|\frac{\partial \phi_T(\zeta|\theta_i^0)}{\partial \theta^{(j)}}\right| \leq \sqrt{s} U_1(\theta_i^0) \|\zeta\|_2. \tag{2.153}
$$

By combining (2.148) and (2.153), we obtain as $\ell \to \infty$

$$
\begin{aligned}
&\int_{a < \|\bar{\zeta}\|_2 < \gamma_1 \sqrt{N_\ell}} A d\bar{\zeta} \\
\leq &\sqrt{N_\ell} \int_{a < \|\bar{\zeta}\|_2 < \gamma_1 \sqrt{N_\ell}} \exp\left(-\frac{N_\ell - 1}{4N_\ell} \bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta}\right) \sqrt{s} U_1(\theta_i^0) \left(\frac{\|\bar{\zeta}\|_2}{\sqrt{N_\ell}}\right) \\
&+ \exp\left(-\frac{1}{2}\bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta}\right) \left(\left|\frac{1}{\sqrt{N_\ell}} \bar{\zeta}^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}}\right| + \left|\frac{1}{2N_\ell} \bar{\zeta}^T \frac{\partial \Lambda_{\theta_i^0}}{\partial \theta^{(j)}} \bar{\zeta}\right|\right) d\bar{\zeta} \\
\leq &\int_{a < \|\bar{\zeta}\|_2 < \gamma_1 \sqrt{N_\ell}} 2 \exp\left(-\frac{1}{8} \bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta}\right) C(\theta_i^0, s) \left(\|\bar{\zeta}\|_2 + \|\bar{\zeta}\|_2^2\right) d\bar{\zeta} \\
\to &C(\theta_i^0, s) \int_{\|\bar{\zeta}\|_2 > a} 2 \exp\left(-\frac{1}{8} \bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta}\right) \left(\|\bar{\zeta}\|_2 + \|\bar{\zeta}\|_2^2\right) d\bar{\zeta},
\end{aligned} \tag{2.154}
$$

where in the second inequality we impose $N_\ell \geq 2$ since it's the limit that is of interest, and $C(\theta_i^0, s)$ is a constant that depends on θ_i^0 and s.

Finally by (2.149) and (2.152), when $\|\bar{\zeta}\|_2 \leq a$

$$
\sqrt{N_\ell} \left(e^{-\frac{\boldsymbol{i}}{\sqrt{N_\ell}} \bar{\zeta}^T \lambda_{\theta_i^0}} \phi_T\left(\frac{\bar{\zeta}}{\sqrt{N_\ell}}|\theta_i^0\right)\right)^{N_\ell - 1} \frac{\partial \phi_T(\frac{\bar{\zeta}}{\sqrt{N_\ell}}|\theta_i^0)}{\partial \theta^{(j)}} \to \exp\left(-\frac{1}{2}\bar{\zeta}^T \Lambda_{\theta_i^0} \bar{\zeta}\right) \boldsymbol{i} \bar{\zeta}^T \frac{\partial \lambda_{\theta_i^0}}{\partial \theta^{(j)}}.
$$

Moreover

$$\sqrt{N_\ell}\exp\left(\frac{\boldsymbol{i}}{\sqrt{N_\ell}}\bar{\zeta}^T\lambda_{\theta_i^0}-\frac{1}{2}\bar{\zeta}^T\Lambda_{\theta_i^0}\bar{\zeta}\right)\left(\frac{\boldsymbol{i}}{\sqrt{N_\ell}}\bar{\zeta}^T\frac{\partial\lambda_{\theta_i^0}}{\partial\theta^{(j)}}-\frac{1}{2N_\ell}\bar{\zeta}^T\frac{\partial\Lambda_{\theta_i^0}}{\partial\theta^{(j)}}\bar{\zeta}\right)$$
$$\to\exp\left(-\frac{1}{2}\bar{\zeta}^T\Lambda_{\theta_i^0}\bar{\zeta}\right)\boldsymbol{i}\bar{\zeta}^T\frac{\partial\lambda_{\theta_i^0}}{\partial\theta^{(j)}}$$

and hence $\lim_{\ell\to\infty}A=0$ when $\|\bar{\zeta}\|_2\le a$. One can also find an integrable envelope function for A when $\|\bar{\zeta}\|_2\le a$ in similar steps as (2.154), and then by Dominated Convergence Theorem,

$$\int_{\|\bar{\zeta}\|_2\le a}Ad\bar{\zeta}\to 0. \tag{2.155}$$

Plug (2.155), (2.154) and (2.151) into (2.150) and (2.147),

$$\begin{aligned}&\limsup_{\ell\to\infty}\ K_\ell(j)\\ \le& C(\theta_i^0,s)\int_{\|\bar{\zeta}\|_2>a}2\exp\left(-\frac{1}{8}\bar{\zeta}^T\Lambda_{\theta_i^0}\bar{\zeta}\right)\left(\|\bar{\zeta}\|_2+\|\bar{\zeta}\|_2^2\right)d\bar{\zeta}.\end{aligned}$$

Letting $a\to\infty$ in the above display yields $K_\ell(j)\to 0$, which together with (2.146) imply $\hat{J}_\ell=o(\sqrt{N_\ell}\|\theta_i^\ell-\eta_i^\ell\|_2)$. □

2.13 Proofs and auxiliary lemmas of Section 2.6

2.13.1 Proof of Theorem 2.6.2

Proof of Theorem 2.6.2:

a) **Step 1** Write n_1 for $n_1(G_0)$ in the proof for clean presentation. Note that (B3) implies that $\theta\mapsto P_\theta$ from $(\Theta,\|\cdot\|_2)$ to $(\{P_\theta\}_{\theta\in\Theta},h)$ is continuous. Then due to Lemma 2.5.5 and Lemma 2.3.2 e), for any $N\ge n_1\vee n_0$,

$$\begin{aligned}h(p_{G,N},p_{G_0,N})\ge V(p_{G,N},p_{G_0,N})&\ge C(G_0,\Theta_1)W_1(G,G_0)\\ &\ge C(G_0,\Theta_1)D_1(G,G_0),\quad\forall G\in\mathcal{E}_{k_0}(\Theta_1).\end{aligned} \tag{2.156}$$

Moreover, by Lemma 2.5.4 for any $N\ge n_1$, and $\forall G\in\mathcal{E}_{k_0}(\Theta_1):W_1(G,G_0)<c(G_0,N)$

$$h(p_{G,N},p_{G_0,N})\ge V(p_{G,N},p_{G_0,N})\ge C(G_0)D_N(G,G_0) \tag{2.157}$$

where $c(G_0,N)$ is a constant that depends on G_0 and N. In the rest of the proof $N\ge n_1\vee n_0$ is implicitly imposed.

By (2.156), for any $\epsilon > 0$,

$$\{G \in \mathcal{E}_{k_0}(\Theta_1) : h(p_{G,N}, p_{G_0,N}) \leq \epsilon\} \subset \left\{G \in \mathcal{E}_{k_0}(\Theta_1) : D_1(G, G_0) \leq \frac{\epsilon}{C(G_0, \Theta_1)}\right\}. \tag{2.158}$$

Recall that $\Theta_1 \subset \Theta \subset \mathbb{R}^q$ and Π is supported on $\mathcal{E}_{k_0}(\Theta_1)$. Then for any $j \in \mathbb{N}$, by (2.158)

$$\begin{aligned}\Pi\left(h(p_{G,N}, p_{G_0,N}) \leq 2j\epsilon\right) &\leq \Pi\left(D_1(G, G_0) \leq \frac{2j\epsilon}{C(G_0, \Theta_1)}\right) \\ &\lesssim k_0! \left(\frac{2j\epsilon}{C(G_0, \Theta_1)}\right)^{k_0 - 1} \left(\frac{2j\epsilon}{C(G_0, \Theta_1)}\right)^{qk_0},\end{aligned} \tag{2.159}$$

where the last inequality follows from (B1).

Using the argument similar to Lemma 3.2(a) in [Ngu16] for any $G = \sum_{i=1}^{k_0} p_i \delta_{\theta_i} \in \mathcal{E}_{k_0}(\Theta_1)$

$$\begin{aligned}K(p_{G_0,N}, p_{G,N}) \leq& NW_{\alpha_0}^{\alpha_0}(G, G_0) \\ \leq& C(\text{diam}(\Theta_1), \alpha_0) N \min_{\tau \in S_{k_0}} \sum_{i=1}^{k_0} \left(\|\theta_{\tau(i)} - \theta_i^0\|_2^{\alpha_0} + |p_{\tau(i)} - p_i^0|\right),\end{aligned}$$

where the first inequality follows from (B3) and the second inequality follows from Lemma 2.3.2 b). Then

$$\Pi\left(K(p_{G_0,N}, p_{G,N}) \leq \epsilon^2\right) \gtrsim \left(\frac{\epsilon^2}{C(\text{diam}(\Theta_1), \alpha_0) N}\right)^{qk_0/\alpha_0} \left(\frac{\epsilon^2}{C(\text{diam}(\Theta_1), \alpha_0) N}\right)^{k_0 - 1}. \tag{2.160}$$

Combine (2.159) and (2.160),

$$\begin{aligned}&\frac{\Pi\left(h(p_{G,N}, p_{G_0,N}) \leq 2j\epsilon\right)}{\Pi\left(K(P_{G_0,N}, P_{G,N}) \leq \epsilon^2\right)} \\ &\qquad \leq C(G_0, \Theta_1, q, \alpha_0, k_0) j^{qk_0 + k_0 - 1} N^{qk_0/\alpha_0 + k_0 - 1} \epsilon^{-qk_0(2/\alpha_0 - 1) - (k_0 - 1)}.\end{aligned}$$

By Remark 2.6.1 $\alpha_0 \leq 2$. Then based on the last display one may verify with

$$\epsilon_{m,N} = C(G_0, \Theta_1, q, k_0, \alpha_0, \beta_0) \sqrt{\frac{\ln(mN)}{m}}$$

for some large enough constant $C(G_0, \Theta_1, q, k_0, \alpha_0, \beta_0)$,

$$\frac{\Pi\left(h(p_{G,N}, p_{G_0,N}) \leq 2j\epsilon_{m,N}\right)}{\Pi\left(K(p_{G_0,N}, p_{G,N}) \leq \epsilon_{m,N}^2\right)} \leq \exp\left(\frac{1}{8} jm\epsilon_{m,N}^2\right).$$

Step 2 By (2.158),

$$\begin{aligned}
&\sup_{\epsilon\geq\epsilon_{m,N}} \ln N\left(\frac{1}{2}\epsilon, \{p_{G,N}: G\in\mathcal{E}_{k_0}(\Theta_1), h(p_{G,N},p_{G_0,N})\leq 2\epsilon\}, h\right)\\
\leq&\sup_{\epsilon\geq\epsilon_{m,N}} \ln N\left(\frac{1}{2}\epsilon, \left\{p_{G,N}: G\in\mathcal{E}_{k_0}(\Theta_1), D_1(G,G_0)\leq \frac{2\epsilon}{C(G_0,\Theta_1)}\right\}, h\right)\\
\leq& qk_0\ln\left(1+\frac{4\times 8^{\frac{1}{\beta_0}}}{C(G_0,\Theta_1)}N^{\frac{1}{2\beta_0}}\epsilon_{m,N}^{-(\frac{1}{\beta_0}-1)}\right)+(k_0-1)\ln\left(1+160\epsilon_{m,N}^{-2}\right),
\end{aligned}$$

where the last inequality follows from Lemma 2.13.1. By Remark 2.6.1 $\beta_0\leq 1$. Then based on the last display one may verify with

$$\epsilon_{m,N}=C(G_0,\Theta_1,q,k_0,\alpha_0,\beta_0)\sqrt{\frac{\ln(mN)}{m}}$$

for some large enough constant $C(G_0,\Theta_1,q,k_0,\alpha_0,\beta_0)$,

$$\sup_{\epsilon\geq\epsilon_{m,N}} \ln N\left(\frac{1}{2}\epsilon, \{p_{G,N}: G\in\mathcal{E}_{k_0}(\Theta_1), h(p_{G,N},p_{G_0,N})\leq 2\epsilon\}, h\right)\leq m\epsilon_{m,N}^2.$$

Step 3 Now we invoke Theorem 8.11 in [GvdV17][5], for every $\bar{M}_m\to\infty$,

$$\Pi(G\in\mathcal{E}_{k_0}(\Theta_1): h(p_{G,N},p_{G_0,N})\geq \bar{M}_m\epsilon_{m,N}|X_{[N]}^1,X_{[N]}^2,\ldots,X_{[N]}^m)\to 0 \tag{2.161}$$

in $\bigotimes^m P_{G_0,N}$-probability as $m\to\infty$ while fixing N. By (2.157) and applying the union bound,

$$\begin{aligned}
&\Pi(G\in\mathcal{E}_{k_0}(\Theta_1): D_N(G,G_0)\geq \frac{\bar{M}_m}{C(G_0)}\epsilon_{m,N}|X_{[N]}^1,X_{[N]}^2,\ldots,X_{[N]}^m)\\
\leq&\Pi(G\in\mathcal{E}_{k_0}(\Theta_1): h(p_{G,N},p_{G_0,N})\geq \bar{M}_m\epsilon_{m,N}|X_{[N]}^1,X_{[N]}^2,\ldots,X_{[N]}^m)\\
&\quad+\Pi(W_1(G,G_0)>c(G_0,N)|X_{[N]}^1,X_{[N]}^2,\ldots,X_{[N]}^m)\\
\to&0
\end{aligned}$$

in $\bigotimes^m P_{G_0,N}$-probability as $m\to\infty$ by (2.161) to the first term. The reason that the second term vanishes is as follows. By (2.156)

$$\{G\in\mathcal{E}_{k_0}(\Theta_1): W_1(G,G_0)>c(G_0,N)\}$$

[5]The Hellinger distance defined in [GvdV17] differs from our definition by a factor of constant. But this constant factor only affect the coefficients of $\epsilon_{m,\bar{N}_m}$ but not the conclusion of convergence.

$$\subset\{G \in \mathcal{E}_{k_0}(\Theta_1) : h(p_{G,N}, p_{G_0,N}) > C(G_0, N, \Theta_1)\}$$

for some constant $C(G_0, N, \Theta_1)$. For some slow-increasing $\bar{M}'_m$ such that $\bar{M}'_m \epsilon_{m,N} \to 0$ as $m \to \infty$,

$$\begin{aligned}&\{G \in \mathcal{E}_{k_0}(\Theta_1) : h(p_{G,N}, p_{G_0,N}) > C(G_0, N, \Theta_1)\}\\ \subset&\{G \in \mathcal{E}_{k_0}(\Theta_1) : h(p_{G,N}, p_{G_0,N}) > \bar{M}'_m \epsilon_{m,N}\}\end{aligned}$$

holds for large m. Combining the last two displays and (2.161) yields as $m \to \infty$

$$\Pi(G \in \mathcal{E}_{k_0}(\Theta_1) : W_1(G, G_0) > c(G_0, N)|X^1_{[N_1]}, \ldots, X^m_{[N_m]}) \to 0.$$

b) If the additional condition of part b) is satisfied, then by Remark 2.5.2 , $n_1(G_0) = 1$. That is, the claim of part a) holds for $n_1(G_0) = 1$.

□

2.13.2 Proof of Theorem 2.6.5

Define the root average square Hellinger metric:

$$d_{m,h}(G, G_0) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} h^2(p_{G,N_i}, p_{G_0,N_i})}.$$

Proof of Theorem 2.6.5: Step 1

a) Write n_1 for $n_1(G_0)$ in the proof for clean presentation. Note that (B3) implies that $\theta \mapsto P_\theta$ from $(\Theta, \|\cdot\|_2)$ to $(\{P_\theta\}_{\theta\in\Theta}, h)$ is continuous. Then due to Lemma 2.5.5 and Lemma 2.3.2 e), for any $N \geq n_1 \vee n_0$, and any $G \in \mathcal{E}_{k_0}(\Theta_1)$,

$$h(p_{G,N}, p_{G_0,N}) \geq V(p_{G,N}, p_{G_0,N}) \geq C(G_0, \Theta_1) W_1(G, G_0) \geq C(G_0, \Theta_1) D_1(G, G_0). \tag{2.162}$$

By (2.162) holds, for all $G \in \mathcal{E}_{k_0}(\Theta_1)$

$$d_{m,h}(G, G_0) \geq C(G_0, \Theta_1) W_1(G, G_0) \geq C(G_0, \Theta_1) D_1(G, G_0). \tag{2.163}$$

Then

$$\{G \in \mathcal{E}_{k_0}(\Theta_1) : d_{m,h}(G, G_0) \leq \epsilon\} \subset \left\{G \in \mathcal{E}_{k_0}(\Theta_1) : D_1(G, G_0) \leq \frac{\epsilon}{C(G_0, \Theta_1)}\right\} \tag{2.164}$$

and thus for any $j \in \mathbb{N}$,

$$\Pi\left(d_{m,h}(G, G_0) \leq 2j\epsilon\right) \leq \Pi\left(D_1(G, G_0) \leq \frac{2j\epsilon}{C(G_0, \Theta_1)}\right) \lesssim k_0! \left(\frac{2j\epsilon}{C(G_0, \Theta_1)}\right)^{k_0-1} \left(\frac{2j\epsilon}{C(G_0, \Theta_1)}\right)^{qk_0}, \quad (2.165)$$

where the last inequality follows from (B1).

By an argument is similar to Lemma 3.2(a) in [Ngu16], for any $G = \sum_{i=1}^{k_0} p_i \delta_{\theta_i} \in \mathcal{E}_{k_0}(\Theta_1)$

$$\begin{aligned} K(p_{G_0,N_i}, p_{G,N_i}) \leq & N_i W_{\alpha_0}^{\alpha_0}(G, G_0) \\ \leq & N_i C(\text{diam}(\Theta_1), \alpha_0) \min_{\tau \in S_{k_0}} \sum_{i=1}^{k_0} \left(\|\theta_{\tau(i)} - \theta_i^0\|_2^{\alpha_0} + |p_{\tau(i)} - p_i^0|\right), \end{aligned}$$

where the second inequality follows from Lemma 2.3.2 b) and (B3). Then

$$\frac{1}{m}\sum_{i=1}^{m} K(p_{G_0,N_i}, p_{G,N_i}) \leq \bar{N}_m C(\text{diam}(\Theta_1), \alpha_0) \min_{\tau \in S_{k_0}} \sum_{i=1}^{k_0} \left(\|\theta_{\tau(i)} - \theta_i^0\|_2^{\alpha_0} + |p_{\tau(i)} - p_i^0|\right),$$

and

$$\begin{aligned} & \Pi\left(\frac{1}{m}\sum_{i=1}^{m} K(p_{G_0,N_i}, p_{G,N_i}) \leq \epsilon^2\right) \\ \gtrsim & \left(\frac{\epsilon^2}{\bar{N}_m C(\text{diam}(\Theta_1), \alpha_0)}\right)^{qk_0/\alpha_0} \left(\frac{\epsilon^2}{\bar{N}_m C(\text{diam}(\Theta_1), \alpha_0)}\right)^{k_0-1}. \end{aligned} \quad (2.166)$$

Combine (2.165) and (2.166),

$$\begin{aligned} & \frac{\Pi\left(d_{m,h}(G, G_0) \leq 2j\epsilon\right)}{\Pi\left(\frac{1}{m}\sum_{i=1}^{m} K(p_{G_0,N_i}, p_{G,N_i}) \leq \epsilon^2\right)} \\ & \leq C(G_0, \Theta_1, q, \alpha_0, k_0) j^{qk_0+k_0-1} \bar{N}_m^{qk_0/\alpha_0+k_0-1} \epsilon^{-qk_0(2/\alpha_0-1)-(k_0-1)}. \end{aligned}$$

Recall by Remark 2.6.1 $\alpha_0 \leq 2$. Then based on the last display one may verify with $\epsilon_{m,\bar{N}_m} = C(G_0, \Theta, q, k_0, \alpha_0, \beta_0)\sqrt{\frac{\ln(m\bar{N}_m)}{m}}$ for some large enough constant $C(G_0, \Theta_1, q, k_0, \alpha_0, \beta_0)$,

$$\frac{\Pi\left(d_{m,h}(G, G_0) \leq 2j\epsilon_{m,\bar{N}_m}\right)}{\Pi\left(\frac{1}{m}\sum_{i=1}^{m} K(p_{G_0,N_i}, p_{G,N_i}) \leq \epsilon_{m,\bar{N}_m}^2\right)} \leq \exp\left(\frac{1}{4} jm\epsilon_{m,\bar{N}_m}^2\right).$$

Step 2 By (2.164),

$$\begin{aligned}
&\sup_{\epsilon\geq\epsilon_{m,\bar{N}_m}} \ln N\left(\frac{1}{36}\epsilon, \{G\in\mathcal{E}_{k_0}(\Theta_1): d_{m,h}(G,G_0)\leq 2\epsilon\}, d_{m,h}\right)\\
\leq &\sup_{\epsilon\geq\epsilon_{m,\bar{N}_m}} \ln N\left(\frac{1}{36}\epsilon, \left\{G\in\mathcal{E}_{k_0}(\Theta_1): D_1(G,G_0)\leq \frac{2\epsilon}{C(G_0,\Theta_1)}\right\}, d_{m,h}\right)\\
\leq & qk_0\ln\left(1+\frac{4\times 144^{\frac{1}{\beta_0}}}{C(G_0,\Theta_1)}\bar{N}_m^{\frac{1}{2\beta_0}}\epsilon_{m,\bar{N}_m}^{-(\frac{1}{\beta_0}-1)}\right)+(k_0-1)\ln\left(1+10\times 72^2\epsilon_{m,\bar{N}_m}^{-2}\right),
\end{aligned}$$

where the last inequality follows from Lemma 2.13.3. Recall by Remark 2.6.1 $\beta_0\leq 1$. Then based on the last display one may verify with $\epsilon_{m,\bar{N}_m}=C(G_0,\Theta_1,q,k_0,\alpha_0,\beta_0)\sqrt{\frac{\ln(m\bar{N}_m)}{m}}$ for some large enough constant $C(G_0,\Theta_1,q,k_0,\alpha_0,\beta_0)$,

$$\sup_{\epsilon\geq\epsilon_{m,\bar{N}_m}} \ln N\left(\frac{1}{36}\epsilon, \{G\in\mathcal{E}_{k_0}(\Theta_1): d_{m,h}(G,G_0)\leq 2\epsilon\}, d_{m,h}\right)\leq m\epsilon_{m,\bar{N}_m}^2. \tag{2.167}$$

Step 3 Now we invoke Theorem 8.23 in [GvdV17][6], we have for every $\bar{M}_m\to\infty$,

$$\Pi(G\in\mathcal{E}_{k_0}(\Theta_1): d_{m,h}(G,G_0)\geq \bar{M}_m\epsilon_{m,\bar{N}_m}|X^1_{[N_1]},\ldots,X^m_{[N_m]})\to 0 \tag{2.168}$$

in $P_{G_0,N_1}\bigotimes\cdots\bigotimes P_{G_0,N_m}$-probability as $m\to\infty$. Since $n_1\leq N_i\leq N_0:=\sup_i N_i$, by Lemma 2.5.4 for $G\in\mathcal{E}_{k_0}(\Theta)$ satisfying $W_1(G,G_0)<c(G_0,N_0)$

$$d_{m,h}(G,G_0)\geq C(G_0)\sqrt{\frac{1}{m}\sum_{i=1}^m D^2_{N_i}(G,G_0)}. \tag{2.169}$$

By Lemma 2.13.4 for $G=\sum_{j=1}^{k_0}p_j\delta_{\theta_j}\in\mathcal{E}_{k_0}(\Theta_1)$ satisfying

$$D_1(G,G_0)<\frac{1}{2}\rho:=\frac{1}{2}\min_{1\leq i<j\leq k_0}\|\theta_i^0-\theta_j^0\|_2,$$

there exists a $\tau\in S_{k_0}$ such that

$$\sqrt{\frac{1}{m}\sum_{i=1}^m D^2_{N_i}(G,G_0)}=\sqrt{\frac{1}{m}\sum_{i=1}^m\left(\sqrt{N_i}\sum_{j=1}^{k_0}\|\theta_{\tau(j)}-\theta_j^0\|_2+\sum_{j=1}^{k_0}|p_{\tau(j)}-p_j^0|\right)^2}$$

[6]The Hellinger distance defined in [GvdV17] differs from our definition by a factor of constant. But this constant factor only affect the coefficients of $\epsilon_{m,\bar{N}_m}$ but not the conclusion of convergence.

$$
\begin{aligned}
&\geq \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(N_i\left(\sum_{j=1}^{k_0}\|\theta_{\tau(j)}-\theta_j^0\|_2\right)^2+\left(\sum_{j=1}^{k_0}|p_{\tau(j)}-p_j^0|\right)^2\right)}\\
&=\sqrt{\bar{N}_m\left(\sum_{j=1}^{k_0}\|\theta_{\tau(j)}-\theta_j^0\|_2\right)^2+\left(\sum_{j=1}^{k_0}|p_{\tau(j)}-p_j^0|\right)^2}\\
&\geq \frac{1}{\sqrt{2}}\left(\sqrt{\bar{N}_m}\sum_{j=1}^{k_0}\|\theta_{\tau(j)}-\theta_j^0\|_2+\sum_{j=1}^{k_0}|p_{\tau(j)}-p_j^0|\right)\\
&=\frac{1}{\sqrt{2}}D_{\bar{N}_m}(G,G_0). \qquad (2.170)
\end{aligned}
$$

Let $\mathcal{G}=\{G\in\mathcal{E}_{k_0}(\Theta_1)|W_1(G,G_0)<c(G_0,N_0),D_1(G,G_0)<\frac{1}{2}\rho\}$. Then combine (2.169) and (2.170), for any $G\in\mathcal{G}$

$$
d_{m,h}(G,G_0)\geq\frac{C(G_0)}{\sqrt{2}}\bar{D}_{\bar{N}_m}(G,G_0).
$$

By the union bound,

$$
\begin{aligned}
&\Pi(G\in\mathcal{E}_{k_0}(\Theta_1):D_{\bar{N}_m}(G,G_0)\geq\frac{\sqrt{2}\bar{M}_m}{C(G_0)}\epsilon_{m,\bar{N}_m}|X^1_{[N_1]},\ldots,X^m_{[N_m]})\\
\leq&\Pi(G\in\mathcal{E}_{k_0}(\Theta_1):d_{m,h}(G,G_0)\geq\bar{M}_m\epsilon_{m,\bar{N}_m}|X^1_{[N_1]},\ldots,X^m_{[N_m]})+\Pi(\mathcal{G}^c|X^1_{[N_1]},\ldots,X^m_{[N_m]})\\
\to&0
\end{aligned}
$$

in $\bigotimes_{i=1}^m P_{G_0,N_i}$-probability as $m\to\infty$ by applying (2.168) to the first term. The reason that the second term vanishes is as follows. By (2.163),

$$
\mathcal{G}^c\subset\{G\in\mathcal{E}_{k_0}(\Theta):d_{m,h}(G,G_0)>C(G_0,\rho,N_0,\Theta_1)\}
$$

for some constant $C(G_0,\rho,N_0,\Theta_1)>0$. For some slow-increasing $\bar{M}'_m$ such that $\bar{M}'_m\epsilon_{m,\bar{N}_m}\to 0$ as $m\to\infty$,

$$
\begin{aligned}
&\{G\in\mathcal{E}_{k_0}(\Theta_1):d_{m,h}(G,G_0)>C(G_0,\rho,N_0,\Theta_1)\}\\
\subset&\{G\in\mathcal{E}_{k_0}(\Theta_1):d_{m,h}(G,G_0)>\bar{M}'_m\epsilon_{m,\bar{N}_m}\}
\end{aligned}
$$

holds for large m. Combining the last two displays and (2.168) yields

$$
\Pi(\mathcal{G}^c|X^1_{[N_1]},\ldots,X^m_{[N_m]})\to 0.
$$

The proof is concluded.

b) If the additional condition of part b) is satisfied, then then by Remark 2.5.2 , $n_1(G_0) = 1$. That is, the claim of part a) holds for $n_1(G_0) = 1$.

□

2.13.3 Auxiliary Lemmas for Section 2.6

For B a subset of metric space with metric D, the minimal number of balls with centers in B and of radius ϵ needed to cover B is known as the ϵ-covering number of B and is denoted by $N(\epsilon, B, D)$.

Lemma 2.13.1. *Fix $G_0 \in \mathcal{E}_{k_0}(\Theta)$. Suppose $h(f(x|\theta_1), f(x|\theta_2)) \leq L_2\|\theta_1 - \theta_2\|_2^{\beta_0}$ for some $0 < \beta_0 \leq 1$ and some $L_2 > 0$ where θ_1, θ_2 are any two distinct elements in Θ.*

$$N\left(\frac{1}{2}\epsilon, \left\{p_{G,N} : G \in \mathcal{E}_{k_0}(\Theta), D_1(G, G_0) \leq \frac{2\epsilon}{C(G_0, \textit{diam}(\Theta))}\right\}, h\right)$$
$$\leq \left(1 + \frac{4 \times 8^{\frac{1}{\beta_0}}}{C(G_0, \textit{diam}(\Theta))} N^{\frac{1}{2\beta_0}} \epsilon^{-(\frac{1}{\beta_0}-1)}\right)^{qk_0} \left(1 + 160\epsilon^{-2}\right)^{k_0 - 1}.$$

Proof: Consider an η_1-net Λ_i with minimum cardinality of $\{\theta : \|\theta - \theta_i^0\|_2 \leq \frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}\}$ and an η_2-net $\bar{\Lambda}$ with minimum cardinality of k_0-probability simplex $\{p \in \mathbb{R}^{k_0} : \sum_{i=1}^{k_0} p_i = 1, p_i \geq 0\}$ under the l_1 distance. Construct a set $\tilde{\Lambda} = \{\tilde{G} = \sum_{i=1}^{k_0} p_i \delta_{\theta_i} : (p_1, \ldots, p_{k_0}) \in \bar{\Lambda}, \theta_i \in \Lambda_i\}$. Then for any $G \in \mathcal{E}_{k_0}(\Theta)$ satisfying $D_1(G, G_0) \leq \frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}$, there exists some $\tilde{G} \in \tilde{\Lambda}$, such that by Lemma 2.7.2

$$h(p_{G,N}, p_{\tilde{G},N}) \leq \sqrt{N}\eta_1^{\beta_0} + \frac{1}{\sqrt{2}}\sqrt{\eta_2}.$$

Thus $\left\{p_{G,N} : G \in \tilde{\Lambda}\right\}$ is a $\left(\sqrt{N}\eta_1^{\beta_0} + \frac{1}{\sqrt{2}}\sqrt{\eta_2}\right)$-net of

$$\left\{p_{G,N} : G \in \mathcal{E}_{k_0}(\Theta), D_1(G, G_0) \leq \frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}\right\}.$$

Since $\tilde{\Lambda}$ is not necessarily subset of $\mathcal{E}_{k_0}(\Theta)$,

$$N\left(2\left(\sqrt{N}\eta_1^{\beta_0} + \frac{1}{\sqrt{2}}\sqrt{\eta_2}\right), \left\{p_{G,N} : G \in \mathcal{E}_{k_0}(\Theta), D_1(G, G_0) \leq \frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}\right\}, h\right)$$
$$\leq |\tilde{\Lambda}| = |\bar{\Lambda}| \prod_{i=1}^{k_0} |\Lambda_i|. \tag{2.171}$$

Now specify $\eta_1 = \left(\frac{\epsilon}{8\sqrt{N}}\right)^{\frac{1}{\beta_0}}$ and thus

$$|\Lambda_i| \leq \left(1 + 2\frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}/\eta_1\right)^q = \left(1 + \frac{4 \times 8^{\frac{1}{\beta_0}}}{C(G_0, \text{diam}(\Theta))} N^{\frac{1}{2\beta_0}} \epsilon^{-(\frac{1}{\beta_0}-1)}\right)^q.$$

Moreover, specify $\eta_2 = \frac{\epsilon^2}{32}$ and by Lemma A.4 of [GVDV01]

$$|\bar{\Lambda}| \leq \left(1 + \frac{5}{\eta_2}\right)^{k_0-1} = \left(1 + 160\epsilon^{-2}\right)^{k_0-1}.$$

Plug the specified η_1 and η_2 into (2.171) and the proof is complete. □

Lemma 2.13.2. *Consider a full rank exponential family's density function $f(x|\theta)$ w.r.t. a dominating measure μ on $\mathfrak{X}$, which takes the form*

$$f(x|\theta) = \exp\left(\theta^T T(x) - A(\theta)\right) h(x),$$

where $\Theta = \{\theta | A(\theta) < \infty\} \subset \mathbb{R}^s$ is the parameter space of θ.

a) For any $\theta_0 \in \Theta^\circ$

$$\limsup_{\theta \to \theta_0} \frac{h(f(x|\theta), f(x|\theta_0))}{\|\theta - \theta_0\|_2} \leq \sqrt{\lambda_{max}(\nabla^2_\theta A(\theta_0))/8},$$

where $\lambda_{max}(\cdot)$ is the maximum eigenvalue of a symmetric matrix.

b) For convex compact subset $\Theta' \subset \Theta^\circ$, there exists $L_2 > 0$ such that

$$h(f(x|\theta_1), f(x|\theta_2)) \leq L_2 \|\theta_1 - \theta_2\|_2 \quad \forall\, \theta_1, \theta_2 \in \Theta'.$$

Proof: a) It is easy to calculate

$$1 - h^2(f(x|\theta_1), f(x|\theta_2)) = \exp\left(A\left(\frac{\theta_1 + \theta_2}{2}\right) - \frac{A(\theta_1) + A(\theta_2)}{2}\right). \tag{2.172}$$

Let $g(\theta) = \exp\left(A\left(\frac{\theta_0+\theta}{2}\right) - \frac{A(\theta_0)+A(\theta)}{2}\right)$. It is easy to verify that $g(\theta_0) = 1$, $\nabla g(\theta_0) = 0$ and $\nabla^2 g(\theta_0) = -\frac{1}{4}\nabla^2 A(\theta_0)$. Then by (2.172)

$$\limsup_{\theta \to \theta_0} \frac{h^2(f(x|\theta), f(x|\theta_0))}{\|\theta - \theta_0\|_2^2} = \limsup_{\theta \to \theta_0} -\frac{g(\theta) - g(\theta_0) - \langle \nabla g(\theta_0), \theta - \theta_0 \rangle}{\|\theta - \theta_0\|_2^2} \tag{2.173}$$

$$
\begin{aligned}
&= \limsup_{\theta \to \theta_0} \frac{\frac{1}{8}(\theta-\theta_0)^T \nabla^2 A(\theta_0)(\theta-\theta_0) + o(\|\theta-\theta_0\|_2^2)}{\|\theta-\theta_0\|_2^2} \\
&\leq \limsup_{\theta \to \theta_0} \left(\frac{1}{8} \lambda_{\max}(\nabla^2 A(\theta_0)) + o(1) \right) \\
&= \frac{1}{8} \lambda_{\max}(\nabla^2 A(\theta_0)).
\end{aligned}
$$

b) For each $\theta, \theta_0 \in \Theta'$, by (2.173),

$$
\begin{aligned}
\frac{h^2(f(x|\theta), f(x|\theta_0))}{\|\theta-\theta_0\|_2^2} &= - \frac{g(\theta) - g(\theta_0) - \langle \nabla g(\theta_0), \theta-\theta_0 \rangle}{\|\theta-\theta_0\|_2^2} \\
&= - \frac{\frac{1}{8}(\theta-\theta_0)^T \nabla^2 g(\xi)(\theta-\theta_0)}{\|\theta-\theta_0\|_2^2} \\
&\leq \frac{1}{8} \sup_{\theta \in \Theta'} \lambda_{\max}(-\nabla^2 g(\theta)),
\end{aligned}
$$

where the second equality follows by Taylor Theorem with ξ in the line joining θ and θ_0 due to the convexity of Θ' and Taylor theorem. The result then follows with $L_2 = \sqrt{\frac{1}{8} \sup_{\theta \in \Theta'} \lambda_{\max}(-\nabla^2 g(\theta))}$, which is finite since $\nabla^2 g(\theta)$, as function of $A(\theta)$ and its gradient and hessian, is continuous on Θ°.

□

Lemma 2.13.3. *Fix $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta)$. Suppose $h(f(x|\theta_1), f(x|\theta_2)) \leq L_2 \|\theta_1 - \theta_2\|_2^{\beta_0}$ for some $0 < \beta_0 \leq 1$ and some $L_2 > 0$ where θ_1, θ_2 are any two distinct elements in Θ.*

$$
\begin{aligned}
&N\left(\frac{1}{36}\epsilon, \left\{ G \in \mathcal{E}_{k_0}(\Theta) : D_1(G, G_0) \leq \frac{2\epsilon}{C(G_0, \textit{diam}(\Theta))} \right\}, d_{m,h} \right) \\
&\leq \left(1 + \frac{4 \times 144^{\frac{1}{\beta_0}}}{C(G_0, \textit{diam}(\Theta))} \bar{N}_m^{\frac{1}{2\beta_0}} \epsilon^{-(\frac{1}{\beta_0}-1)} \right)^{qk_0} \left(1 + 10 \times 72^2 \epsilon^{-2} \right)^{k_0 - 1}.
\end{aligned}
$$

Proof: Consider an η_1-net Λ_i with minimum cardinality of $\{\theta : \|\theta - \theta_i^0\|_2 \leq \frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}\}$ and an η_2-net $\bar{\Lambda}$ with minimum cardinality of k_0-probability simplex $\{p \in \mathbb{R}^{k_0} : \sum_{i=1}^{k_0} p_i = 1, p_i \geq 0\}$ under the l_1 distance. Construct a set $\tilde{\Lambda} = \{\tilde{G} = \sum_{i=1}^{k_0} p_i \delta_{\theta_i} : (p_1, \ldots, p_{k_0}) \in \bar{\Lambda}, \theta_i \in \Lambda_i\}$. Then for any $G \in \mathcal{E}_{k_0}(\Theta)$ satisfying $D_1(G, G_0) \leq \frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}$, there exists some $\tilde{G} \in \tilde{\Lambda}$, such that by Lemma 2.7.2

$$
h^2(p_{G,N_i}, p_{\tilde{G},N_i}) \leq \left(\sqrt{N_i} \eta_1^{\beta_0} + \frac{1}{\sqrt{2}} \sqrt{\eta_2} \right)^2 \leq 2 \left(N_i \eta_1^{2\beta_0} + \frac{1}{2} \eta_2 \right).
$$

Thus

$$
d_{m,h}(G, \tilde{G}) \leq \sqrt{2 \bar{N}_m \eta_1^{2\beta_0} + \eta_2}.
$$

As a result $\tilde{\Lambda}$ is a $\sqrt{2\bar{N}_m\eta_1^{2\beta_0}+\eta_2}$-net of $\left\{G \in \mathcal{E}_{k_0}(\Theta) : D_1(G, G_0) \leq \frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}\right\}$. Since $\tilde{\Lambda}$ is not necessarily subset of $\mathcal{E}_{k_0}(\Theta)$,

$$N\left(2\sqrt{2\bar{N}_m\eta_1^{2\beta_0}+\eta_2}, \left\{G \in \mathcal{E}_{k_0}(\Theta) : D_1(G, G_0) \leq \frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}\right\}, d_{m,h}\right) \leq |\tilde{\Lambda}|$$
$$= |\bar{\Lambda}| \prod_{i=1}^{k_0} |\Lambda_i|. \quad (2.174)$$

Now specify $\eta_1 = \left(\frac{\epsilon}{144\sqrt{\bar{N}_m}}\right)^{\frac{1}{\beta_0}}$ and thus

$$|\Lambda_i| \leq \left(1 + 2\frac{2\epsilon}{C(G_0, \text{diam}(\Theta))}/\eta_1\right)^q = \left(1 + \frac{4 \times 144^{\frac{1}{\beta_0}}}{C(G_0, \text{diam}(\Theta))}\bar{N}_m^{\frac{1}{2\beta_0}}\epsilon^{-(\frac{1}{\beta_0}-1)}\right)^q.$$

Moreover, specify $\eta_2 = \frac{1}{2}\left(\frac{\epsilon}{72}\right)^2$ and by Lemma A.4 of [GVDV01] $|\bar{\Lambda}| \leq \left(1 + \frac{5}{\eta_2}\right)^{k_0-1} = (1 + 10 \times 72^2\epsilon^{-2})^{k_0-1}$. Plug the specified η_1 and η_2 into (2.174) and the proof is complete. □

Lemma 2.13.4. *For $G_0 = \sum_{i=1}^{k_0} p_i\delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta)$ with $\rho = \min_{1\leq i<j\leq k_0} \|\theta_i^0 - \theta_j^0\|_2$. If $G = \sum_{i=1}^{k_0} p_i\delta_{\theta_i} \in \mathcal{E}_{k_0}(\Theta)$ satisfying $D_1(G, G_0) < \frac{1}{2}\rho$, then there exists a unique $\tau \in S_{k_0}$ such that for all real number $r \geq 1$*

$$D_r(G, G_0) = \sum_{i=1}^{k_0} \left(\sqrt{r}\|\theta_{\tau(i)} - \theta_i^0\|_2 + |p_{\tau(i)} - p_i^0|\right).$$

Proof: Let τ be any one in S_{k_0} such that

$$D_1(G, G_0) = \sum_{i=1}^{k_0} \left(\|\theta_{\tau(i)} - \theta_i^0\|_2 + |p_{\tau(i)} - p_i^0|\right).$$

For any $j \neq \tau(i)$, $\|\theta_j - \theta_i^0\|_2 \geq \|\theta_{\tau^{-1}(j)}^0 - \theta_i^0\|_2 - \|\theta_j - \theta_{\tau^{-1}(j)}^0\|_2 > \rho - \rho/2 = \frac{\rho}{2}$. Then for any $\tau' \in S_{k_0}$ that is not τ and for any real number $r \geq 1$

$$\sum_{i=1}^{k_0} \left(\sqrt{r}\|\theta_{\tau'(i)} - \theta_i^0\|_2 + |p_{\tau'(i)} - p_i^0|\right) > \sqrt{r}\frac{\rho}{2} > \sqrt{r}D_1(G, G_0)$$
$$\geq \sum_{i=1}^{k_0} \left(\sqrt{r}\|\theta_{\tau(i)} - \theta_i^0\|_2 + |p_{\tau(i)} - p_i^0|\right),$$

which with $r = 1$ shows our choice of τ is unique and with $r \geq 1$ shows τ is the optimal permutation

for $D_r(G, G_0)$. $\square$

2.14 Proofs in Section 2.7

Proof of Lemma 2.7.2: Step 1: Suppose $p'_i = p_i$ for any $i \in [k_0]$. In this case,

$$\begin{aligned} h^2(P_{G,N}, P_{G',N}) =& h^2\left(\sum_{i=1}^{k_0} p_i P_{\theta_i,N}, \sum_{i=1}^{k_0} p_i P_{\theta'_i,N}\right) \\ \leq& \sum_{i=1}^{k_0} p_i h^2\left(P_{\theta_i,N}, P_{\theta'_i,N}\right) \\ \leq& N \sum_{i=1}^{k_0} p_i h^2\left(P_{\theta_i}, P_{\theta'_i}\right) \\ \leq& N \max_{1\leq i\leq k_0} h^2\left(P_{\theta_i}, P_{\theta'_i}\right), \end{aligned}$$

where the first inequality follows from the joint convexity of any f-divergences (of which squared Hellinger distance is a member), and the second inequality follows from

$$h^2\left(P_{\theta_i,N}, P_{\theta'_i,N}\right) = 1 - \left(1 - h^2\left(P_{\theta_i}, P_{\theta'_i}\right)\right)^N \leq N h^2\left(P_{\theta_i}, P_{\theta'_i}\right).$$

Step 2: Suppose $\theta'_i = \theta_i$ for any $i \in [k_0]$. Let $\boldsymbol{p} = (p_1, p_2, \ldots, p_{k_0})$ be the discrete probability measure associated to the weights of G and define $\boldsymbol{p}'$ similarly. Consider any $Q = (q_{ij})_{i,j=1}^{k_0}$ to be a coupling of $\boldsymbol{p}$ and $\boldsymbol{p}'$. Then

$$\begin{aligned} h^2(P_{G,N}, P_{G',N}) =& h^2\left(\sum_{i=1}^{k_0}\sum_{j=1}^{k_0} q_{ij} P_{\theta_i,N}, \sum_{i=1}^{k_0}\sum_{j=1}^{k_0} q_{ij} P_{\theta_j,N}\right) \\ \leq& \sum_{i=1}^{k_0}\sum_{j=1}^{k_0} q_{ij} h^2\left(P_{\theta_i,N}, P_{\theta_j,N}\right) \\ \leq& \sum_{i=1}^{k_0}\sum_{j=1}^{k_0} q_{ij} \mathbf{1}(\theta_i \neq \theta_j), \end{aligned} \tag{2.175}$$

where the first inequality follows from the joint convexity of any f-divergence, and the second inequality follow from the hellinger distance is upper bounded by 1. Since (2.175) holds for any coupling Q of $\boldsymbol{p}$ and $\boldsymbol{p}'$,

$$h^2(P_{G,N}, P_{G',N}) \leq \inf_Q \sum_{i=1}^{k_0}\sum_{j=1}^{k_0} q_{ij} \mathbf{1}(\theta_i \neq \theta_j) = V(\boldsymbol{p}, \boldsymbol{p}') = \frac{1}{2}\sum_{i=1}^{k_0} |p_i - p'_i|.$$

Step 3: General case. Let $G'' = \sum_{i=1}^{k_0} p_i \delta_{\theta'_i}$. Then by triangular inequality, Step 1 and Step 2,

$$h(P_{G,N}, P_{G',N}) \le h(P_{G,N}, P_{G'',N}) + h(P_{G'',N}, P_{G',N}) \le \sqrt{N} \max_{1 \le i \le k_0} h\left(P_{\theta_i}, P_{\theta'_i}\right) + \sqrt{\frac{1}{2} \sum_{i=1}^{k_0} |p_i - p'_i|}.$$

Finally, notice the above procedure does not depend on the specific order of atoms of G and G', and thus the proof is complete. □

Proof of Lemma 2.7.3: Since $\liminf_{\theta \to \theta_j^0} \frac{h(P_\theta, P_{\theta_j^0})}{\|\theta - \theta_j^0\|_2} < \infty$, there exists a sequences $\{\theta_j^k\}_{k=1}^\infty \subset \Theta \backslash \cup_{i=1}^{k_0} \{\theta_i^0\}$ such that $\theta_j^k \to \theta_j^0$ and

$$h(P_{\theta_j^k}, P_{\theta_j^0}) \le \gamma \|\theta_j^k - \theta_j^0\|_2 \tag{2.176}$$

for some $\gamma \in (0, \infty)$. Suppose

$$\limsup_{N \to \infty} \liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{h(P_{G,N}, P_{G_0,N})}{D_{\psi(N)}(G, G_0)} = \beta \in (0, \infty],$$

then there exists subsequences $N_\ell \to \infty$ such that for any ℓ

$$\liminf_{\substack{G \overset{W_1}{\to} G_0 \\ G \in \mathcal{E}_{k_0}(\Theta)}} \frac{h(P_{G,N_\ell}, P_{G_0,N_\ell})}{D_{\psi(N_\ell)}(G, G_0)} \ge \frac{3}{4}\beta.$$

Thus for each ℓ, there exists $\theta_j^{k_\ell}$ such that $G_\ell = p_j^0 \delta_{\theta_j^{k_\ell}} + \sum_{i=1, i \ne j}^{k_0} p_i^0 \delta_{\theta_i^0} \in \mathcal{E}_{k_0}(\Theta) \backslash \{G_0\}$, and

$$\frac{h(P_{G_\ell,N_\ell}, P_{G_0,N_\ell})}{D_{\psi(N_\ell)}(G_\ell, G_0)} \ge \frac{\beta}{2}.$$

By our choice of G_ℓ, for sufficiently large ℓ

$$h(P_{G_\ell,N_\ell}, P_{G_0,N_\ell}) \ge \frac{\beta}{2} D_{\psi(N_\ell)}(G_\ell, G_0) = \frac{\beta}{2} \sqrt{\psi(N_\ell)} \|\theta_j^{k_\ell} - \theta_j^0\|_2.$$

On the other hand, by Lemma 2.7.2,

$$h(P_{G_\ell,N_\ell}, P_{G_0,N_\ell}) \le \sqrt{N_\ell} h(P_{\theta_j^{k_\ell}}, P_{\theta_j^0}).$$

Combining the last two displays,

$$\frac{\beta}{2} \leq \sqrt{\frac{N_\ell}{\psi(N_\ell)}} \frac{h(P_{\theta_j^{k_\ell}}, P_{\theta_j^0})}{\|\theta_j^{k_\ell} - \theta_j^0\|_2} \leq \gamma \sqrt{\frac{N_\ell}{\psi(N_\ell)}} \to 0, \quad \text{as } \ell \to \infty,$$

where the second inequality follows from (2.176). But the last display contradicts with $\beta \in (0, \infty]$. □

Proof of Theorem 2.7.5: a) Choose a set of distinct $k_0 - 1$ points $\{\theta_i\}_{i=1}^{k_0-1} \subset \Theta \backslash \{\theta_0\}$ satisfying

$$\rho_1 = \min_{0 \leq i < j \leq k_0 - 1} h(P_{\theta_i}, P_{\theta_j}) > 0.$$

Let $\rho = \min_{0 \leq i < j \leq k_0 - 1} \|\theta_i - \theta_j\|_2$. Since $\limsup_{\theta \to \theta_0} \frac{h(P_\theta, P_{\theta_0})}{\|\theta - \theta_0\|_2^{\beta_0}} < \infty$, there exist $\gamma \in (0, \infty)$ and $r_0 \in (0, \min\{\rho, (\rho_1/\gamma)^{1/\beta_0}\})$ such that

$$\frac{h(P_\theta, P_{\theta_0})}{\|\theta - \theta_0\|_2^{\beta_0}} < \gamma, \quad \forall 0 < \|\theta - \theta_0\|_2 < r_0. \tag{2.177}$$

Consider $G_1 = \sum_{i=1}^{k_0} \frac{1}{k_0} \delta_{\theta_i^1} \in \mathcal{E}_{k_0}(\Theta)$ and $G_2 = \sum_{i=1}^{k_0} \frac{1}{k_0} \delta_{\theta_i^2} \in \mathcal{E}_{k_0}(\Theta)$ with $\theta_i^1 = \theta_i^2 = \theta_i \in \Theta \backslash \{\theta_0\}$ for $i \in [k_0 - 1]$ and $\theta_{k_0}^1 = \theta_0$, $\theta_{k_0}^2 = \theta$ satisfying $\|\theta - \theta_0\|_2 = 2\epsilon < r_0$. Here $\epsilon \in (0, r_0/2)$ is a constant to be determined. Then $d_{\boldsymbol{\theta}}(G_1, G_2) = 2\epsilon$. Moreover, $h(P_\theta, P_{\theta_0}) \leq \gamma (2\epsilon)^{\beta_0} < \rho_1$.

By two-point Le Cam bound (i.e. (15.14) in [Wai19])

$$\inf_{\hat{G} \in \mathcal{E}_{k_0}(\Theta)} \sup_{G \in \mathcal{E}_{k_0}(\Theta)} \mathbb{E}_{\bigotimes^m P_{G,N}} d_{\boldsymbol{\theta}}(G, \hat{G}) \geq \frac{\epsilon}{2} \left(1 - V\left(\bigotimes^m P_{G_1,N}, \bigotimes^m P_{G_2,N}\right)\right). \tag{2.178}$$

Notice

$$V\left(\bigotimes^m P_{G_1,N}, \bigotimes^m P_{G_2,N}\right) \leq h\left(\bigotimes^m P_{G_1,N}, \bigotimes^m P_{G_2,N}\right) \leq \sqrt{m} h\left(P_{G_1,N}, P_{G_2,N}\right).$$

With our choice of G_1 and G_2, by Lemma 2.7.2, the last display becomes

$$\begin{aligned} V\left(\bigotimes^m P_{G_1,N}, \bigotimes^m P_{G_2,N}\right) \leq & \sqrt{m}\sqrt{N} \min_{\tau \in S_{k_0}} \max_{1 \leq i \leq k_0} h\left(P_{\theta_i^1}, P_{\theta_{\tau(i)}^2}\right) \\ = & \sqrt{m}\sqrt{N} h\left(P_{\theta_0}, P_\theta\right) \\ \leq & \sqrt{m}\sqrt{N} \gamma (2\epsilon)^{\beta_0}, \end{aligned} \tag{2.179}$$

where the equality follows from

$$\min_{\tau\in S_{k_0}}\max_{1\le i\le k_0} h(P_{\theta_i^1},P_{\theta^2_{\tau(i)}})=h(P_{\theta_{k_0}^1},P_{\theta^2_{k_0}})=h\left(P_{\theta_0},P_\theta\right)$$

due to $h\left(P_{\theta_0},P_\theta\right)<\rho_1$. Plug (2.179) into (2.178),

$$\inf_{\hat{G}\in\mathcal{E}_{k_0}(\Theta)}\sup_{G\in\mathcal{E}_{k_0}(\Theta)}\mathbb{E}_{\bigotimes^m P_{G,N}}d_{\boldsymbol{\theta}}(G,\hat{G})\ge\frac{\epsilon}{2}(1-\gamma\sqrt{m}\sqrt{N}(2\epsilon)^{\beta_0}). \tag{2.180}$$

Consider any $a\in(0,1)$ satisfying $a>1-\gamma r_0^{\beta_0}$ and let $2\epsilon=\left(\frac{1-a}{\gamma\sqrt{m}\sqrt{N}}\right)^{\frac{1}{\beta_0}}$. Then $2\epsilon\in(0,r_0)$. Plug the specified ϵ into (2.180), then the right hand side in the above display becomes

$$\frac{a}{4}\left(\frac{1-a}{\gamma\sqrt{m}\sqrt{N}}\right)^{\frac{1}{\beta_0}}=C(\beta_0)\left(\frac{1}{\sqrt{m}\sqrt{N}}\right)^{\frac{1}{\beta_0}},$$

where $C(\beta_0)$ depends on β_0. Notice a,γ,r_0 are just some absolute constants that depends on the probability family $\{P_\theta\}_{\theta\in\Theta}$.

b) Consider $k_0>3$. Let $0<\epsilon<(\frac{1}{3}-\frac{1}{3(k_0-2)})/2$. Consider $G_1=\sum_{i=1}^2\frac{1}{3}\delta_{\theta_i}+\sum_{i=3}^{k_0}\frac{1}{3(k_0-2)}\delta_{\theta_i}\in\mathcal{E}_{k_0}(\Theta)$ and $G_2=(\frac{1}{3}-\epsilon)\delta_{\theta_1}+(\frac{1}{3}+\epsilon)\delta_{\theta_2}+\sum_{i=3}^{k_0}\frac{1}{3(k_0-2)}\delta_{\theta_i}\in\mathcal{E}_{k_0}(\Theta)$. By the range of ϵ, $G_2\in\mathcal{E}_{k_0}(\Theta)$ and $d_{\boldsymbol{p}}(G_1,G_2)=2\epsilon$. Similar to the proof of a),

$$\inf_{\hat{G}\in\mathcal{E}_{k_0}(\Theta)}\sup_{G\in\mathcal{E}_{k_0}(\Theta)}\mathbb{E}_{\bigotimes^m P_{G,N}}d_{\boldsymbol{p}}(\hat{G},G)\ge\frac{\epsilon}{2}\left(1-\sqrt{m}h\left(P_{G_1,N},P_{G_2,N}\right)\right).$$

With our choice of G_1 and G_2, by Lemma 2.7.2,

$$h\left(P_{G_1,N},P_{G_2,N}\right)\le\sqrt{\frac{1}{2}\times 2\epsilon}=\sqrt{\epsilon}.$$

Combine the last two displays,

$$\inf_{\hat{G}\in\mathcal{E}_{k_0}(\Theta)}\sup_{G\in\mathcal{E}_{k_0}(\Theta)}\mathbb{E}_{\bigotimes^m P_{G,N}}d_{\boldsymbol{p}}(\hat{G},G)\ge\frac{\epsilon}{2}\left(1-\sqrt{m}\sqrt{\epsilon}\right).$$

The proof is complete by specifying $\epsilon=\frac{1}{m}(\frac{1}{3}-\frac{1}{3(k_0-2)})/4<(\frac{1}{3}-\frac{1}{3(k_0-2)})/2$. The case for $k_0=2$ and $k_0=3$ follows similarly.

c) The proof follows from a), b) and (2.38).

□

2.15 Proofs in Section 2.8

Proof of Lemma 2.8.2: Let $c(\cdot)$ be a positive constant that depends on its parameters in this proof.

Claim 1: There exists $c_6 > 0$ that depends only on $d, j_1, \ldots, j_d$ such that

$$S_{\min}(A(x)) \geq c_6 |x|^{-(j_d - j_1)(d-1)}$$

for any $|x| > 1$. Suppose this is not true, then there is $\{x_m\}_{m=1}^{\infty}$ such that $|x_m| > 1$, and

$$|x_m|^{(j_d - j_1)(d-1)} S_{\min}(A(x_m)) \to 0. \tag{2.181}$$

Let $B(x,t) = |x|^t A(x)$ with t being some positive number to be specified. The characteristic polynomial of $B(x,t)B^T(x,t)$ is

$$\det\left(\lambda I - B(x,t)B^T(x,t)\right) = \lambda^d + \sum_{i=0}^{d-1} \gamma_i(x,t)\lambda^i.$$

When $|x| > 1$, since $|A_{\alpha\beta}(x)| \leq c_4(d, j_1, \cdots, j_d)|x|^{j_d - j_1}$ for any $\alpha, \beta \in [d]$, the entries of $B(x,t)B^T(x,t)$ are bounded by $d\left(c_4(d, j_1, \cdots, j_d)|x|^{(j_d - j_1 + t)}\right)^2$. Thus

$$|\gamma_i(x,t)| \leq c_8(d, j_1, \cdots, j_d)\left(|x|^{(j_d - j_1 + t)}\right)^{2(d-i)}$$

for $1 \leq i \leq d-1$. Moreover,

$$|\gamma_0(x,t)| = \left|x^{dt}\det(A(x))\right|^2 = (\prod_{i=1}^{d} j_i!)^2 |x|^{2dt} = c_5(d, j_1, \cdots, j_d)|x|^{2dt},$$

with $c_5(d, j_1, \cdots, j_d) = (\prod_{i=1}^{d} j_i!)^2 > 0$. Let $\lambda_{\min}(x,t) \geq 0$ be the smallest eigenvalue of $B(x,t)B^T(x,t)$. Then

$$\lambda_{\min}^d(x,t) + \sum_{i=0}^{d-1} \gamma_i(x,t)\lambda_{\min}^i(x,t) = 0.$$

When $x \neq 0$, $\lambda_{\min}(x,t) > 0$ since $\gamma_0(x,t) \neq 0$. Thus when $x \neq 0$,

$$\frac{1}{\lambda_{\min}(x,t)} = -\frac{1}{\gamma_0(x,t)}\lambda_{\min}^{d-1}(x,t) - \sum_{i=1}^{d-1} \frac{\gamma_i(x,t)}{\gamma_0(x,t)}\lambda_{\min}^{i-1}(x,t). \tag{2.182}$$

Moreover, when $|x| > 1$, $\left|\frac{\gamma_i(x,t)}{\gamma_0(x,t)}\right| \leq \frac{c_8(d,j_1,\cdots,j_d)}{c_5(d,j_1,\cdots,j_d)} \frac{|x|^{2(j_d - j_1)(d-i)}}{|x|^{2ti}} \leq \frac{c_8(d,j_1,\cdots,j_d)}{c_5(d,j_1,\cdots,j_d)} \frac{|x|^{2(j_d - j_1)(d-1)}}{|x|^{2t}}$ for any

$1 \le i \le d-1$. Then by (2.181), $\lambda_{\min}(x_m, t_0) = \left(|x_m|^{(j_d - j_1)(d-1)} S_{\min}(A(x_m))\right)^2 \to 0$, where $t_0 = (j_d - j_1)(d-1)$, so $\frac{1}{\lambda_{\min}(x,t_0)} \to \infty$. On the other hand, since $|\frac{1}{\gamma_0(x_m,t_0)}|$ and $\frac{\gamma_i(x_m,t_0)}{\gamma_0(x_m,t_0)}$ are bounded and $\lambda_{\min}(x_m, t_0) \to 0$,

$$\limsup_{m\to\infty} \left| -\frac{1}{\gamma_0(x_m,t_0)} \lambda_{\min}^{d-1}(x_m,t_0) - \sum_{i=1}^{d-1} \frac{\gamma_i(x_m,t_0)}{\gamma_0(x_m,t_0)} \lambda_{\min}^{i-1}(x_m,t_0) \right| = \limsup_{n\to\infty} \left| \frac{\gamma_1(x_m,\alpha_0)}{\gamma_0(x_m,\alpha_0)} \right| \le \frac{c_8(d,j_1,\cdots,j_d)}{c_5(d,j_1,\cdots,j_d)}.$$

These contradict with (2.182) and hence the claim at the beginning of this paragraph is established.

Since $S_{\min}(A(x)) > 0$ on $|x| \le 1$, as a continuous function on a compact set,

$$\min_{x\in[-1,1]} S_{\min}(A(x)) \ge c_7 > 0.$$

Then take $c_3 = \min\{c_6, c_7\}$ and the proof is complete. □

Proofs of Lemma 2.8.3: Let $\psi_w(x) = w^T T x$. Then

$$\left(\frac{d^{j_1}\psi_w(x)}{dx^{j_1}}, \frac{d^{j_2}\psi_w(x)}{dx^{j_2}}, \ldots, \frac{d^{j_d}\psi_w(x)}{dx^{j_d}} \right)^T = A(x)w$$

where $A(x) \in \mathbb{R}^{d\times d}$ with entries $A_{\alpha\beta}(x) = 0$ for $\alpha > \beta$ and $A_{\alpha\beta}(x) = \frac{j_\beta!}{(j_\beta - j_\alpha)!} x^{j_\beta - j_\alpha}$ for $\alpha \le \beta$. Then for any $w \in S^{d-1}$,

$$\max_{1\le i\le d} \left| \frac{d^{j_i}\psi_w(x)}{dx^{j_i}} \right| = \|A(x)w\|_\infty \ge \frac{1}{\sqrt{d}} \|A(x)w\|_2 \ge \frac{1}{\sqrt{d}} S_{\min}(A(x)) \ge \frac{1}{\sqrt{d}} c_3 \max\{1, |x|\}^{-\alpha_0}, \tag{2.183}$$

where $\alpha_0 = (j_d - j_1)(d-1)$, $S_{\min}(A(x))$ is the smallest singular value of $A(x)$ and the last inequality follows from Lemma 2.8.2.

Case 1: $j_1 > 1$.

Partition the real line according to the increasing sequence $\{a_t\}_{t=-\infty}^{\infty}$ where

$$a_t = \begin{cases} 2a_{t+1} & t \le -1 \\ \lfloor -c_1 \rfloor - 1 & t = 0 \\ b_t & 1 \le t \le \ell \\ \lceil c_1 \rceil + 1 & t = \ell + 1 \\ 2a_{t-1} & t \ge \ell + 2 \end{cases}.$$

For $t \leq -1$, by (2.183) we know $\max\limits_{1\leq i\leq d}\left|\frac{d^{j_i}\psi_w(x)}{dx^{j_i}}\right| \geq \frac{1}{\sqrt{d}}c_3|a_t|^{-\alpha_0}$ for all $x \in [a_t, a_{t+1}]$. In order to appeal to Lemma 2.8.1, we need to specify the points $\{t_\beta\}_{\beta=0}^{\beta_0}$ with $t_0 = a_t < t_1 < \ldots < t_{\beta_0} = a_{t+1}$, where $\{t_\beta\}_{\beta=1}^{\beta_0-1}$ is defined as the set of roots in (a_t, a_{t+1}) of any of the following $d-1$ equations,

$$\left|\frac{d^{j_i}\psi_w(x)}{dx^{j_i}}\right| = \frac{1}{\sqrt{d}}c_3|a_t|^{-\alpha_0}, \quad i \in [d-1].$$

Thus $\{t_\beta\}_{\beta=0}^{\beta_0}$ is a partition of $[a_t, a_{t+1}]$ such that for each $0 \leq \beta \leq \beta_0 - 1$, $\left|\frac{d^{j_{k_\beta}}\psi_w(x)}{dx^{j_{k_\beta}}}\right| \geq \frac{1}{\sqrt{d}}c_3|a_i|^{-\alpha_0}$ holds for some index $k_\beta \in [d]$ and for all $x \in [t_\beta, t_{\beta+1}]$. Since $\frac{d^{j_m}\psi_w(x)}{dx^{j_m}}$ is polynomial of degree $j_d - j_m$, it follows that $\beta_0 - 1 \leq 2\sum_{m=1}^d (j_d - j_m)$. Let $\tilde{c}_0$ be the maximum of $\{\tilde{c}_{j_m}\}_{m=1}^d$, where $\tilde{c}_{j_m}$ are the coefficients c_k corresponds to $k = j_m$ in Lemma 2.8.1. Then by Lemma 2.8.1, for $\lambda > 1$

$$\begin{aligned}
&\left|\int_{[t_\beta, t_{\beta+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)}f(x)dx\right| \\
\leq&\tilde{c}_0\left(\frac{c_3|a_t|^{-\alpha_0}\lambda}{\sqrt{d}}\right)^{-\frac{1}{j_{k_\beta}}}\left(|f(t_{\beta+1})| + \int_{[t_\beta, t_{\beta+1}]}|f'(x)|dx\right) \\
\leq&\tilde{c}_0\max\left\{c_3^{-\frac{1}{j_1}}, c_3^{-\frac{1}{j_d}}\right\}(\sqrt{d})^{\frac{1}{j_1}}\lambda^{-\frac{1}{j_d}}(|a_t|^{\alpha_0})^{\frac{1}{j_1}}\left(f(a_{t+1}) + \int_{[t_\beta, t_{\beta+1}]}|f'(x)|dx\right), \qquad (2.184)
\end{aligned}$$

where the last step follows from $f(x)$ being increasing on $(-\infty, -c_1)$. Then for $\lambda > 1$

$$\begin{aligned}
&\left|\int_{[a_t, a_{t+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)}f(x)dx\right| \\
\leq&\sum_{\beta=0}^{\beta_0-1}\left|\int_{[t_\beta, t_{\beta+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)}f(x)dx\right| \\
\overset{(*)}{\leq}&\tilde{c}_0\max\left\{c_3^{-\frac{1}{j_1}}, c_3^{-\frac{1}{j_d}}\right\}(\sqrt{d})^{\frac{1}{j_1}}\lambda^{-\frac{1}{j_d}}(|a_t|^{\alpha_0})^{\frac{1}{j_1}}\left(\beta_0 f(a_{t+1}) + \int_{[a_t, a_{t+1}]}|f'(x)|dx\right) \\
\overset{(**)}{\leq}&\tilde{c}_0\max\left\{c_3^{-\frac{1}{j_1}}, c_3^{-\frac{1}{j_d}}\right\}(\sqrt{d})^{\frac{1}{j_1}}\lambda^{-\frac{1}{j_d}}2^{\frac{\alpha_0}{j_1}}\left(\beta_1|a_{t+1}|^{\frac{\alpha_0}{j_1}}f(a_{t+1}) + |a_{t+1}|^{\frac{\alpha_0}{j_1}}\int_{[a_t, a_{t+1}]}|f'(x)|dx\right) \\
\leq&C(d, j_1, \cdots, j_d)\lambda^{-\frac{1}{j_d}}\left(|a_{t+1}|^{\frac{\alpha_0}{j_1}}f(a_{t+1}) + \int_{[a_t, a_{t+1}]}|x|^{\frac{\alpha_0}{j_1}}|f'(x)|dx\right), \qquad (2.185)
\end{aligned}$$

where the step $(*)$ follows from (2.184), the step $(**)$ follows from $a_t = 2a_{t+1}$ and $\beta_0 \leq \beta_1 := 2\sum_{m=1}^d (j_d - j_m) + 1$, and the last step follows from $\beta_1 \geq 1$, $|a_t| \geq |x| \geq |a_{t+1}|$ for all $x \in [a_t, a_{t+1}]$ and $C(d, j_1, \cdots, j_d) = \tilde{c}_0\max\left\{c_3^{-\frac{1}{j_1}}, c_3^{-\frac{1}{j_d}}\right\}(\sqrt{d})^{\frac{1}{j_1}}2^{\frac{\alpha_0}{j_1}}\beta_1$.

For $t \geq \ell + 1$, following similar steps as the case $t \leq -1$, one obtain

$$\left| \int_{[a_t, a_{t+1}]} e^{\boldsymbol{i}\lambda\phi_w(x)} f(x) dx \right| \leq C(d, j_1, \cdots, j_d) \lambda^{-\frac{1}{j_d}} \left(|a_t|^{\frac{\alpha_0}{j_1}} f(a_t) + \int_{[a_t, a_{t+1}]} |x|^{\frac{\alpha_0}{j_1}} |f'(x)| dx \right), \tag{2.186}$$

where $C(d, j_1, \cdots, j_d)$ is the same as in (2.185).

For $0 \leq t \leq \ell$, since f' is continuous on (a_t, a_{t+1}) and f' is Lebesgue integrable on $[a_t, a_{t+1}]$, $\lim_{x \to a_{t+1}^-} f(x)$ and $\lim_{x \to a_t^+} f(x)$ exist. Define

$$\tilde{f}(x) = f(x) \mathbf{1}_{(a_t, a_{t+1})}(x) + \mathbf{1}_{\{a_{t+1}\}}(x) \lim_{x \to a_{t+1}^-} f(x) + \mathbf{1}_{\{a_t\}}(x) \lim_{x \to a_t^+} f(x).$$

Then $\tilde{f}(x)$ is absolute continuous on $[a_t, a_{t+1}]$. Moreover, by (2.183) we know $\max_{1 \leq i \leq d} \left| \frac{d^{j_i} \psi_w(x)}{dx^{j_i}} \right| \geq \frac{1}{\sqrt{d}} c_3 (c_1 + 2)^{-\alpha_0}$ for all $x \in [a_t, a_{t+1}]$. Following the same argument as in the previous case, let $\{\tilde{t}_\beta\}_{\beta=0}^{\tilde{\beta}_0}$ with $\tilde{t}_0 = a_t < \tilde{t}_1 < \ldots < \tilde{t}_{\beta_0} = a_{t+1}$, where $\{\tilde{t}_\beta\}_{\beta=1}^{\tilde{\beta}_0 - 1}$ is the set of roots in (a_t, a_{t+1}) of the following $d - 1$ equations

$$\left| \frac{d^{j_i} \psi_w(x)}{dx^{j_i}} \right| = \frac{1}{\sqrt{d}} c_3 (c_1 + 2)^{-\alpha_0}, \quad i \in [d-1].$$

Then $\{\tilde{t}_\beta\}_{\beta=0}^{\tilde{\beta}_0}$ is a partition of $[a_t, a_{t+1}]$ such that for each $0 \leq \beta \leq \tilde{\beta}_0 - 1$, $\left| \frac{d^{j_{k_\beta}} \psi_w(x)}{dx^{j_{k_\beta}}} \right| \geq \frac{1}{\sqrt{d}} c_3 (c_1 + 2)^{-\alpha_0}$ for some $k_\beta \in [d]$ and for all $x \in [\tilde{t}_\beta, \tilde{t}_{\beta+1}]$. Since $\frac{d^{j_m} \psi_w(x)}{dx^{j_m}}$ are polynomial of degree $j_d - j_m$, we have $\tilde{\beta}_0 - 1 \leq 2 \sum_{m=1}^d (j_d - j_m)$. Thus by Lemma 2.8.1, for any $\lambda > 1$

$$\begin{aligned}
&\left| \int_{[\tilde{t}_\beta, \tilde{t}_{\beta+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)} f(x) dx \right| \\
=&\left| \int_{[\tilde{t}_\beta, \tilde{t}_{\beta+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)} \tilde{f}(x) dx \right| \\
\leq& \tilde{c}_0 \left(\frac{c_3 (c_1 + 2)^{-\alpha_0} \lambda}{\sqrt{d}} \right)^{-\frac{1}{j_{k_\beta}}} \left(|\tilde{f}(\tilde{t}_{\beta+1})| + \int_{[\tilde{t}_\beta, \tilde{t}_{\beta+1}]} |f'(x)| dx \right) \\
\leq& \tilde{c}_0 \max \left\{ c_3^{-\frac{1}{j_1}}, c_3^{-\frac{1}{j_d}} \right\} (\sqrt{d})^{\frac{1}{j_1}} \lambda^{-\frac{1}{j_d}} ((c_1 + 2)^{\alpha_0})^{\frac{1}{j_1}} \left(\|f\|_{L^\infty} + \int_{[\tilde{t}_\beta, \tilde{t}_{\beta+1}]} |f'(x)| dx \right),
\end{aligned} \tag{2.187}$$

where the last step follows from $|\tilde{f}(\tilde{t}_{\beta+1})| \leq \|f\|_{L^\infty}$. Then for any $\lambda > 1$

$$\left| \int_{[a_t, a_{t+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)} f(x) dx \right|$$

$$
\begin{aligned}
&\leq \sum_{\beta=0}^{\tilde{\beta}_0-1} \left| \int_{[\tilde{t}_\beta, \tilde{t}_{\beta+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)} f(x)dx \right| \\
&\leq \tilde{c}_0 \max\left\{ c_3^{-\frac{1}{j_1}}, c_3^{-\frac{1}{j_d}} \right\} (\sqrt{d})^{\frac{1}{j_1}} \lambda^{-\frac{1}{j_d}} ((c_1+2)^{\alpha_0})^{\frac{1}{j_1}} \left(\tilde{\beta}_0 \|f\|_{L^\infty} + \int_{[a_t, a_{t+1}]} |f'(x)|dx \right) \\
&\leq C(d, j_1, \ldots, j_d) \lambda^{-\frac{1}{j_d}} (c_1+2)^{\frac{\alpha_0}{j_1}} \left(\|f\|_{L^\infty} + \int_{[a_t, a_{t+1}]} |f'(x)|dx \right), \qquad (2.188)
\end{aligned}
$$

where $C(d, j_1, \cdots, j_d)$ is the same as in (2.185).

Hence,

$$
\begin{aligned}
&\left| \int_{\mathbb{R}} e^{\boldsymbol{i}\lambda\psi_w(x)} f(x)dx \right| \\
&= \left| \sum_{t=-\infty}^{\infty} \int_{[a_t, a_{t+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)} f(x)dx \right| \\
&\leq \sum_{t=-\infty}^{\infty} \left| \int_{[a_t, a_{t+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)} f(x)dx \right| \\
&\overset{(*)}{\leq} C(d, j_1, \ldots, j_d) \lambda^{-\frac{1}{j_d}} (c_1+2)^{\frac{\alpha_0}{j_1}} \\
&\quad \left(\sum_{t\leq -1} |a_{t+1}|^{\frac{\alpha_0}{j_1}} f(a_{t+1}) + \sum_{t\geq \ell+1} |a_t|^{\frac{\alpha_0}{j_1}} f(a_t) + (\ell+1)\|f\|_{L^\infty} + \left\| \left(|x|^{\frac{\alpha_0}{j_1}} + 1 \right) f'(x) \right\|_{L^1} \right) \\
&\overset{(**)}{\leq} C(d, j_1, \ldots, j_d) \lambda^{-\frac{1}{j_d}} (c_1+2)^{\frac{\alpha_0}{j_1}} \\
&\quad \left(\int_{(\infty, -c_1]} |x|^{\frac{\alpha_0}{j_1}} f(x)dx + \int_{[c_1, \infty)} |x|^{\frac{\alpha_0}{j_1}} f(x)dx + (\ell+1)\|f\|_{L^\infty} + \left\| \left(|x|^{\frac{\alpha_0}{j_1}} + 1 \right) f'(x) \right\|_{L^1} \right) \\
&\leq C(d, j_1, \ldots, j_d) \lambda^{-\frac{1}{j_d}} (c_1+2)^{\frac{\alpha_0}{j_1}} \left(\left\| |x|^{\frac{\alpha_0}{j_1}} f(x) \right\|_{L^1} + (\ell+1)\|f\|_{L^\infty} + \left\| \left(|x|^{\frac{\alpha_0}{j_1}} + 1 \right) f'(x) \right\|_{L^1} \right)
\end{aligned} \qquad (2.189)
$$

where the first equality follows from the dominated convergence theorem, the step $(*)$ follows from (2.185), (2.186), (2.188), and the step $(**)$ follows from the monotonicity of $|x|^{\frac{\alpha_0}{j_1}} f$ when $x < -c_1$, $x > c_1$.

Case 2: $j_1 = 1$.

Fix $\forall w \in S^{d-1}$, $\exists x_1 < x_2 < \ldots < x_s$ partition $\mathbb{R}$ into $s+1$ disjoint open intervals such that $\frac{d\psi_w(x)}{dx}$ is monotone on each of those interval. Notice $s \leq j_d - 2$ since $\frac{d\psi_w(x)}{dx}$ is a polynomial of degree $j_d - 1$, and $x_1, x_2, \ldots, x_s$ depend on w. For $t \leq -1$, on $[a_t, a_{t+1}]$ when we subdivide the interval, besides the partition points $\{t_\beta\}_{\beta=0}^{\beta_0}$, $\{x_1, x_2, \ldots, x_s\} \cap [a_t, a_{t+1}]$ should also be added into the partition points. The new partition points set has at most $\beta_0 + 1 + s \leq \beta_1 + j_d$ points and hence subdivide $[a_t, a_{t+1}]$ into at most $\beta_1 + j_d - 1$ intervals such that on each subinterval

$\max_{1\le i\le d}\left|\frac{d^{j_i}\psi_w(x)}{dx^{j_i}}\right| \ge \frac{1}{\sqrt{d}}c_3|a_t|^{-\alpha_0}$ and $\frac{d\psi_w(x)}{dx}$ is monotone. Hence Lemma 2.8.1 (part ii)) can be applied on each subinterval. The rest of steps proceed similarly as in Case 1, and one will obtain

$$\left|\int_{[a_t,a_{t+1}]} e^{\boldsymbol{i}\lambda\psi_w(x)}f(x)dx\right| \le \tilde{C}(d,j_1,\cdots,j_d)\lambda^{-\frac{1}{j_d}}\left(|a_{t+1}|^{\frac{\alpha_0}{j_1}}f(a_{t+1}) + \int_{[a_t,a_{t+1}]}|x|^{\frac{\alpha_0}{j_1}}|f'(x)|dx\right), \tag{2.190}$$

where $\tilde{C}(d,j_1,\cdots,j_d) = \tilde{c}_0\max\left\{c_3^{-\frac{1}{j_1}}, c_3^{-\frac{1}{j_d}}\right\}(\sqrt{d})^{\frac{1}{j_1}}2^{\frac{\alpha_0}{j_1}}(\beta_1+j_d-1)$, a constant that depends only on $d, j_1, \ldots, j_d$. For the same reasoning one can obtain (2.186) for $t \ge \ell+1$ and (2.188) for $0 \le t \le \ell$, both with $C(d,j_1,\cdots,j_d)$ replaced by $\tilde{C}(d,j_1,\cdots,j_d)$. As a result, similar to (2.189),

$$\begin{aligned}&\left|\int_{\mathbb{R}} e^{\boldsymbol{i}\lambda\psi_w(x)}f(x)dx\right| \\ \le&\tilde{C}(d,j_1,\ldots,j_d)\lambda^{-\frac{1}{j_d}}(c_1+2)^{\frac{\alpha_0}{j_1}}\left(\left\||x|^{\frac{\alpha_0}{j_1}}f(x)\right\|_{L^1} + (\ell+1)\|f\|_{L^\infty} + \left\|\left(|x|^{\frac{\alpha_0}{j_1}}+1\right)f'(x)\right\|_{L^1}\right).\end{aligned}$$

□

Proof: By Lemma 2.8.3, when $\|\zeta\|_2 > 1$,

$$|g(\zeta)|^r \le C(f,d,j_1,\ldots,j_d)\|\zeta\|_2^{-\frac{r}{j_d}}.$$

where

$$\begin{aligned}&C(f,r,d,j_1,\ldots,j_d) = \\ &\quad C^r(d,j_1,\ldots,j_d)(c_1+2)^{\alpha_1 r}\left(\||x|^{\alpha_1}f(x)\|_{L^1} + (\ell+1)\|f\|_{L^\infty} + \|(|x|^{\alpha_1}+1)f'(x)\|_{L^1}\right)^r.\end{aligned}$$

Then

$$\begin{aligned}&\int_{\|\zeta\|_2>1}|g(\zeta)|^r d\zeta \\ \le&C(f,r,d,j_1,\ldots,j_d)\int_{\|\zeta\|_2>1}\|\zeta\|_2^{-\frac{r}{j_d}}d\zeta \\ \le&C(f,r,d,j_1,\ldots,j_d)|S^{d-1}|\int_{(1,\infty)}\lambda^{-\frac{r}{j_d}}\lambda^{d-1}d\lambda \\ =&C(r,d,j_1,\ldots,j_d)(c_1+2)^{\alpha_1 r}\left(\||x|^{\alpha_1}f(x)\|_{L^1} + (\ell+1)\|f\|_{L^\infty} + \|(|x|^{\alpha_1}+1)f'(x)\|_{L^1}\right)^r.\end{aligned} \tag{2.191}$$

where the last inequality follows from $\int_{(1,\infty)}\lambda^{-\frac{r}{j_d}}\lambda^{d-1}d\lambda$ is a finite constant that depends on d and

j_d for $r > dj_d$ and $C(r, d, j_1, \ldots, j_d) = C^r(d, j_1, \ldots, j_d)|S^{d-1}| \int_{(1,\infty)} \lambda^{-\frac{r}{j_d}} \lambda^{d-1} d\lambda$. In addition,

$$\int_{\|\zeta\|_2 \leq 1} |g(\zeta)|^r d\zeta \leq \int_{\|\zeta\|_2 \leq 1} \|f\|_{L^1}^r d\zeta = C(d)\|f\|_{L^1}^r, \tag{2.192}$$

where $C(d)$ is a constant that depends on d.

The proof is then completed by combining (2.191) and (2.192) and $(a^r + b^r) \leq (a + b)^r$ for any $a, b > 0, r \geq 1$. □

2.16 Auxiliary lemmas for Section 2.12.2

Consider a family of probabilities $\{P_\theta\}_{\theta \in \Theta}$ on $\mathbb{R}^d$, where θ is the parameter of the family and $\Theta \subset \mathbb{R}^q$ is the parameter space. $\mathbb{E}_\theta$ denotes the expectation under the probability measure P_θ. Consider $\{X_i\}_{i=1}^\infty$ a sequence of independent and identically distributed random vectors from P_{θ_0}. Suppose $\mathbb{E}_{\theta_0} X_1$ exists and define $Z_N = \frac{\sum_{i=1}^N X_i - N\mathbb{E}_{\theta_0} X_1}{\sqrt{N}}$. The next result establishes the density of Z_N converges to that of a multivariate normal distribution.

Lemma 2.16.1 (Local Central Limit Theorem)**.** *Suppose $\{X_i\}_{i=1}^\infty$ a sequence of independent and identically distributed random vectors from P_{θ_0}. Suppose $\mathbb{E}_{\theta_0} X_1$ and $\Lambda_{\theta_0} := \mathbb{E}_{\theta_0}(X_1 - \mathbb{E}_{\theta_0} X_1)(X_1 - \mathbb{E}_{\theta_0} X_1)^T$ exist and Λ_{θ_0} is positive definite. Let the characteristic function of P_θ be $\phi(\zeta|\theta) := \mathbb{E}_\theta e^{i\zeta^T X}$ and suppose there exists $r \geq 1$ such that $|\phi(\zeta|\theta_0)|^r$ is Lebesgue integrable on $\mathbb{R}^d$. Then when $N \geq r$, Z_N has density with respect to Lebesgue measure on $\mathbb{R}^d$, and its density $f_Z(z|\theta_0, N)$ as N tends to infinity converges uniformly in z to $f_{\mathcal{N}}(z|\theta_0)$, the density of $\mathcal{N}(\mathbf{0}, \Lambda_{\theta_0})$, the multivariate normal with mean $\mathbf{0}$ and covariance matrix Λ_{θ_0}.*

The special case for $d = 1$ of the above lemma is Theorem 2 in Section 5, Chapter XV of [Fel08]. That proof generalize to $d > 1$ without much difficulties.

The next lemma is a generalization of the corollary to Lemma 1 in [Sta65].

Lemma 2.16.2. *Consider a random vector $X \in \mathbb{R}^d$ with $\phi(\zeta)$ its characteristic function. Suppose X has density $f(x)$ w.r.t. Lebesgue measure upper bounded by U, and has positive definite covariance matrix Λ. Then for all $\zeta \in \mathbb{R}^d$*

$$|\phi(\zeta)| \leq \exp\left(-\frac{C(d)\|\zeta\|_2^2}{(\|\zeta\|_2^2 \lambda_{\max}(\Lambda) + 1)\lambda_{\max}^{d-1}(\Lambda) U^2}\right),$$

where $C(d)$ is some constant that depends only on d, and $\lambda_{\max}(\Lambda)$ is the largest eigenvalue of Λ.

Proof: It suffices to prove for $\zeta \neq \mathbf{0} \in \mathbb{R}^d$.

Step 1

In this step we prove the special case $\zeta = te_1$ for $t > 0$, where e_1 is the standard basis in $\mathbb{R}^d$.

Define $I(\zeta) = \frac{1}{2}\left(1 - |\phi(\zeta)|^2\right)$ and it is easy to verify

$$|\phi(\zeta)| \leq \exp(-I(\zeta)). \tag{2.193}$$

Denote by $\tilde{f}$ to be the density w.r.t. Lebesgue measure of symmetrized random vector $X - X'$, where X' is an independent copy of X. Then $\tilde{f}$ also has upper bound U and $|\phi(\zeta)|^2$ is the characteristic function of $X - X'$ and

$$|\phi(\zeta)|^2 = \int_{\mathbb{R}^d} e^{i\zeta^T x}\tilde{f}(x)dx = \int_{\mathbb{R}^d} \cos(\zeta^T x)\tilde{f}(x)dx. \tag{2.194}$$

Write $x = (x^{(1)}, \ldots, x^{(d)})$ and let $G_j = \{x \in \mathbb{R}^d | x^{(1)} \in (\frac{j}{t} - \frac{1}{2t}, \frac{j}{t} + \frac{1}{2t}]\}$ be the strip of length $\frac{1}{t}$ centered at $\frac{j}{t}$ across the $x^{(1)}$-axis. Then by (2.194)

$$\begin{aligned}
I(2\pi\zeta) =& \int_{\mathbb{R}^d} \sin^2(\pi\zeta^T x)\tilde{f}(x)dx \\
\geq& \int_B \sin^2(\pi t x^{(1)})\tilde{f}(x)dx \\
=& \sum_{j=-\infty}^{\infty} \int_{G_j \bigcap B} \sin^2(\pi t x^{(1)})\tilde{f}(x)dx \\
=& \sum_{j=-\infty}^{\infty} \int_{G_j \bigcap B} \sin^2(\pi t (x^{(1)} - j/t))\tilde{f}(x)dx \\
\geq& 4t^2 \sum_{j=-\infty}^{\infty} \int_{G_j \bigcap B} (x^{(1)} - j/t)^2 \tilde{f}(x)dx,
\end{aligned} \tag{2.195}$$

where the first inequality follows from $\zeta = te_1$ and B is a subset in $\mathbb{R}^d$ to be determined, and the last inequality follows from $|\sin(\pi x)| \geq 2|x|$ for $|x| \leq \frac{1}{2}$.

Let $B = \{z \in \mathbb{R}^d | |z^{(i)}| < 2\sqrt{d\lambda_{\max}(\Lambda)}\ \forall i \geq 2, \text{ and } |z^{(1)}| < \frac{r}{t} + \frac{1}{2t}\}$ with $r = \min\{b \text{ interger} : \frac{b}{t} + \frac{1}{2t} \geq 2\sqrt{d\lambda_{\max}(\Lambda)}\}$. Then $B \subset \bigcup_{j=-r}^{r} G_j$ and thus (2.195) become

$$\begin{aligned}
I(2\pi\zeta) \geq& 4t^2 \sum_{j=-r}^{r} \int_{G_j \bigcap B} (x^{(1)} - j/t)^2 \tilde{f}(x)dx \\
\overset{(*)}{=}& 4t^2 \sum_{j=-r}^{r} \int_{G} (x^{(1)} - j/t)^2 \tilde{f}(x)\mathbf{1}_{G_j \bigcap B}(x)dx \\
\overset{(**)}{\geq}& 4t^2 \sum_{j=-r}^{r} \frac{Q_j^3}{12U^2(4\sqrt{d\lambda_{\max}(\Lambda)})^{2(d-1)}}
\end{aligned}$$

$$\overset{(***)}{\geq} 4t^2 \frac{Q^3}{12(2r+1)^2U^2(4\sqrt{d\lambda_{\max}(\Lambda)})^{2(d-1)}}, \tag{2.196}$$

where in step $(*)$ $G = \{z \in \mathbb{R}^d || z^{(i)}| < 2\sqrt{d\lambda_{\max}(\Lambda)}\, \forall i \geq 2\}$, step $(**)$ with $Q_j = \int_{G_j \bigcap B} \tilde{f}(x)dx$ follows from Lemma 2.16.3 b) and step $(***)$ with $Q = \sum_{j=-r}^{r} Q_j = \int_B \tilde{f}(x)dx$ follows from Jensen's inequality. (The inequalities in step $(**)$ and $(***)$ are attained with $\tilde{f}(x) = U \sum_{j=-r}^{r} \mathbf{1}_{W_j}(x)$ $a.e.$ $x \in G$ where $W_j = \{z||z^{(i)}| < 2\sqrt{d\lambda_{\max}(\Lambda)}, \forall i \geq 2, \text{ and } |z^{(1)} - j/t| < a\}$ for positive a satisfies
$(2a)(4\sqrt{d\lambda_{\max}(\Lambda)})^{d-1}U(2r+1) = Q.$)

Observe $\{z \in \mathbb{R}^d | z^T(2\Lambda)^{-1}z < 2d\} \subset B$ and then

$$Q = P(X - X' \in B) \geq 1 - P((X - X')^T(2\Lambda)^{-1}(X - X') \geq 2d) \geq \frac{1}{2},$$

where the last step follows from Markov inequality. Moreover by our choice of r, $2r + 1 \leq 4t\sqrt{d\lambda_{\max}(\Lambda)} + 2$. Then (2.196) become

$$\begin{aligned} I(2\pi\zeta) \geq & t^2 \frac{1}{24(4t\sqrt{d\lambda_{\max}(\Lambda)} + 2)^2(4\sqrt{d\lambda_{\max}(\Lambda)})^{2(d-1)}U^2} \\ \geq & \frac{C(d)t^2}{(t^2\lambda_{\max}(\Lambda) + 1)\lambda_{\max}^{d-1}(\Lambda)U^2}, \end{aligned}$$

where $C(d)$ is a constant that depends only on d. The last display replacing $2\pi\zeta = 2\pi t e_1$ by $\zeta = te_1$, together with (2.193) yield the desired conclusion.

Step 2

For any $\zeta \neq 0$, denote $t = \|\zeta\|_2$ and $u_1 = \zeta/\|\zeta\|_2$. Consider an orthogonal matrix U_ζ with its first row u_1^T. Then $\phi(\zeta) = \mathbb{E}e^{itu_1^T X} = \mathbb{E}e^{ite_1^T Z}$ where $Z = U_\zeta X$. Since Z has density $f_Z(z) = f(U_\zeta^T z)$ w.r.t. Lebesgue measure, $f_Z(z)$ has the same upper bound U and positive definite covariance matrix $U_\zeta \Lambda U_\zeta^T$ with the same largest eigenvalue as Λ. The result then follows by applying **Step 1** to $\left|\mathbb{E}e^{ite_1^T Z}\right|$. □

Lemma 2.16.3. *a) Consider a Lebesgue measurable function on $\mathbb{R}$ satisfies $0 \leq f(x) \leq U$ and $\int_{\mathbb{R}} f(x)dx = E \in (0, \infty)$. Then for any $b > 0$*

$$\int_{\mathbb{R}} (x - b)^2 f(x)dx \geq \frac{E^3}{12U^2},$$

and the equality holds if and only if $f(x) = U\mathbf{1}_{[b - \frac{E}{2U}, b + \frac{E}{2U}]}(x)$ $a.e.$.

b) For $a > 0$ define a set $G = \{z \in \mathbb{R}^d || z^{(i)}| < a \quad \forall i \geq 2\}$. Consider a Lebesgue measurable function on $\mathbb{R}^d$ satisfies $0 \leq f(x) \leq U$ on G and $\int_G f(x)dx = E \in (0, \infty)$. Then for any

$b > 0$

$$\int_G (x^{(1)} - b)^2 f(x)dx \geq \frac{E^3}{12U^2(2a)^{2(d-1)}},$$

and the equality holds if and only if $f(x) = U\mathbf{1}_{G_1}(x)$ *a.e.* $x \in G$ *where* $G_1 = [b - \frac{E}{2U(2a)^{d-1}}, b + \frac{E}{2U(2a)^{d-1}}] \times (-a, a)^{d-1}$.

Proof: a) It suffices to prove $b = 0$ since one can do the translation $x' = x - b$ to reduce the general case b to the special case $b = 0$. Let $f_1(x) = f(x)\mathbf{1}_{[-\frac{E}{2U}, \frac{E}{2U}]}(x)$, $f_2(x) = f(x)\mathbf{1}_{[-\frac{E}{2U}, \frac{E}{2U}]^c}(x)$ and $f_U(x) = U\mathbf{1}_{[-\frac{E}{2U}, \frac{E}{2U}]}(x) - f_1(x)$. Then

$$\int_{[-\frac{E}{2U}, \frac{E}{2U}]} f_U(x)dx = E - \int_{[-\frac{E}{2U}, \frac{E}{2U}]} f_1(x)dx = \int_{[-\frac{E}{2U}, \frac{E}{2U}]^c} f_2(x)dx$$

and hence

$$\begin{aligned}
\int_{\mathbb{R}} x^2 f(x)dx =& \int_{[-\frac{E}{2U}, \frac{E}{2U}]} x^2 f_1(x)dx + \int_{[-\frac{E}{2U}, \frac{E}{2U}]^c} x^2 f_2(x)dx \\
\geq& \int_{[-\frac{E}{2U}, \frac{E}{2U}]} x^2 f_1(x)dx + \left(\frac{E}{2U}\right)^2 \int_{[-\frac{E}{2U}, \frac{E}{2U}]^c} f_2(x)dx \\
=& \int_{[-\frac{E}{2U}, \frac{E}{2U}]} x^2 f_1(x)dx + \left(\frac{E}{2U}\right)^2 \int_{[-\frac{E}{2U}, \frac{E}{2U}]} f_U(x)dx \\
\geq& \int_{[-\frac{E}{2U}, \frac{E}{2U}]} x^2 f_1(x)dx + \int_{[-\frac{E}{2U}, \frac{E}{2U}]} x^2 f_U(x)dx \\
=& \int_{[-\frac{E}{2U}, \frac{E}{2U}]} x^2 U dx \\
=& \frac{E^3}{12U^2}.
\end{aligned}$$

The equality holds if and only if the last two inequalities are attained, if and only if $f(x) = U\mathbf{1}_{[-\frac{E}{2U}, \frac{E}{2U}]}(x)$ *a.e.*.

b) It suffices to prove $b = 0$ since one can always do the translation $y^{(1)} = x^{(1)} - b$ and $y^{(i)} = x^{(i)}$ for all $2 \leq i \leq d$ to reduce the general case b to the special case $b = 0$. By Tonelli's Theorem, $h(x^{(1)}) = \int_{(-a,a)^{d-1}} f(x)dx^{(2)} \dots dx^{(d)}$ exists for *a.e.* $x^{(1)}$ and $\int_{\mathbb{R}} h(x^{(1)})dx^{(1)} = E$. Moreover $0 \leq h(x^{(1)}) \leq U(2a)^{d-1}$ *a.e.* . Then by Tonelli's Theorem and a)

$$\int_G (x^{(1)})^2 f(x)dx = \int_{\mathbb{R}} (x^{(1)})^2 h(x^{(1)})dx^{(1)} \geq \frac{E^3}{12U^2(2a)^{2(d-1)}}.$$

The equality holds if and only if $h(x^{(1)}) = U(2a)^{d-1}\mathbf{1}_{[-\frac{E}{2U(2a)^{d-1}}, \frac{E}{2U(2a)^{d-1}}]}(x^{(1)})$ *a.e.*, if and

only if $f(x) = U$ $a.e. x \in [-\frac{E}{2U(2a)^{d-1}}, \frac{E}{2U(2a)^{d-1}}] \times (-a, a)^{d-1}$.

□

CHAPTER 3

Screening in High Dimensional Data

3.1 Introduction

This chapter considers the problem of screening n independent and identically distributed p-variate samples for variables that have high correlation or high partial correlation with at least one other variable in the ultra-high dimensional regime when the sample size $n \leq C_0 \ln p$.[1] In the screening framework one applies a threshold to the sample correlation matrix or the sample partial correlation matrix to detect variables with at least one significant correlation, with the threshold aiming to separate signal from noise. Correlation and partial correlation screening in ultra-high dimensions have become increasingly important in many modern applications as the per-sample cost of collecting high dimensional data is much more costly than per-variable cost. For example, in biomedical settings the cost of high throughput technology, like oligonucleotide gene microchips and RNAseq assays is decreasing, while the cost of biological samples is not decreasing at the same rate [HR15b]. In such situations p is much larger than n.

The ultra-high dimensional regime when $n \leq C_0 \ln p$ is very challenging since the number of samples is insufficient to apply many (if not most) reliable statistical methods. For example, one way to undertake partial correlation screening is to first estimate the population covariance matrix, then obtain the inverse, from which a partial correlation matrix can be estimated. However, to get a reliable estimate of a general covariance matrix, the number of samples n must be at least $O(p)$ as shown in Section 5.4.3. in [Ver12]. Even if the covariance matrix has a special structure like sparsity, covariance estimation requires a number of samples of order $O(\ln p)$ [RBLZ08]. The reader is referred to [DR17, LW18, KOR15, CKG19] and the references therein for recent work in modern high dimensional covariance selection and estimation.

While estimating the covariance matrix or partial correlation matrix is challenging in ultra-high dimensions, recent work has shown that it is possible to accurately test the number of highly (partial) correlated variables under a false positive probability; in particular the probability that a variable

[1] Here C_0 is some universal constant satisfying $C_0 \geq 1$. A “universal constant" or “absolute constant", is a constant that does not depend on any model parameter.

is highly (partially) correlated with at least one other variable [HR11, HR12]. While correlation screening finds variables that have a high marginal correlation with at least one other variable, partial correlation screening identifies variables that have high conditional correlations with one other variable conditioned on the rest. In [HR11], the ultra-high dimensional correlation screening problem is studied under a row-sparse assumption on the population covariance matrix. A phase transition in the number of false positive correlations was mathematically characterized as a function of the correlation threshold and the true covariance. In the case of block sparse covariance, the critical phase transition threshold becomes independent of the true covariance. In [HR12] the partial correlation screening problem was studied, and similar phase transition results as in correlation screening [HR11] were obtained under the block-sparse assumption on the population covariance matrix. The survey [HR15a] reviews the correlation and partial correlation screening problem.

Despite these important advances in correlation and partial correlation screening, the screening framework proposed in [HR11, HR12] has some serious methodological, theoretical and practical shortcomings. For instance, results for partial correlation imposes a highly restrictive block sparsity assumption on the true underlying correlation matrix. The block sparsity in [HR12] assumes only a small group of the variables are allowed to have correlation within the blocks and no correlations with variables outside the block. This assumption is severely restrictive for most modern applications since it is possible for variables to have correlations within a group and also correlations with variables outside their respective groups. Furthermore, expressions for false probabilities in [HR11, HR12] require estimating dependence functionals. Estimating such functionals lead to computationally prohibitive non-parametric estimation, rendering the screening methodology disconnected from the very setting it was designed for.

In this chapter we propose a novel unifying framework for correlation and partial correlation screening that delivers a practical and scalable methodology in the ultra-high dimensional regime, which is simultaneously armed with theoretical safeguards. By making novel and insightful connections to random geometric graphs we demonstrate that the distribution of the number of discoveries tends to a compound Poisson limit. To the best of our knowledge, such a novel limit has not previously appeared in the correlation screening setting. Furthermore, our results are proved in much generality by relaxing block-sparse assumption to a weaker (τ, κ) sparsity assumption, defined in Section 3.2.3, on the population covariance matrix. The block-sparse assumption is a special case of the (τ, κ) sparsity assumption. Resulting approximations under this generalized covariance structure (τ, κ) also do not depend on dependence measures/functionals. The results in this chapter hold for both correlation and partial correlation screening. This important duality naturally stems from new results relating the score representation for correlation screening to that of partial correlation screening. The proofs of the generalized results in this chapter are self-contained and are based on Stein's approximation, concentration of random matrices, and concentrations in

high dimensional balls and spheres.

The theory in this chapter is relevant to hypothesis testing based on the degree distribution of a correlation graph, a problem arising in graph mining, network science, social science, and the natural science [CF06, KC14, Kol09]. Variables having strong sample correlations will appear in the correlation graph as vertices having positive vertex degree. As one sweeps over fixed degree values, the number of such vertices specifies the degree distribution of the graph. From this perspective, this chapter provides a non-asymptotic compound Poisson characterization of the degree distribution for large correlation graphs under relaxed sparsity conditions on the population covariance.

The remainder of this chapter is organized as follows. We begin in Section 3.2 by giving the framework and presenting our main theorem on characterization of the limiting distribution. In Section 3.3 a non-asymptotic version of the main theorem is presented, based on which the main theorem follows. Section 3.4 is devoted to convergence of moments. Section 3.5 provide an extensive study on computing and approximating the parameters in the main theorems. A number of technical proofs and auxiliary results are given in the Appendix.

Notation $\|\cdot\|_2$ for a vector represents its Euclidean distance to the origin. C and c denotes positive universal constants that might defer from line to line. C and c with subscripts are positive finite constants depending only on the parameter in their subscripts and may differ from line to line.

3.2 A unified theorem

3.2.1 Framework

Available is a matrix of multivariate samples

$$\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(n)}]^T = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}, \tag{3.1}$$

where $\{\boldsymbol{x}^{(i)}\}_{i=1}^n \subset \mathbb{R}^p$ are samples from a p-dimensional distribution.

The results in this chapter apply when the $n \times p$ data matrix $\boldsymbol{X}$ follows a vector elliptically contoured distribution. A random matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is called vector-elliptical[2] with positive definite dispersion parameter $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and location parameter $\boldsymbol{\mu}$ if its density satisfies

$$f_{\boldsymbol{X}}(\boldsymbol{X}) = \det(\boldsymbol{\Sigma})^{-n/2}\theta(\mathrm{tr}((\boldsymbol{X} - \mathbf{1}\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}^T - \mathbf{1}\boldsymbol{\mu}^T))), \tag{3.2}$$

for a shaping function $\theta : \mathbb{R} \to \mathbb{R}^+$ such that $\int f_{\boldsymbol{X}}(\boldsymbol{X}) = 1$. In (3.2), $\mathbf{1}$ is a vector with all elements equal to 1, $\mathrm{tr}(\cdot)$ is the trace of a matrix and $\det(\cdot)$ is the determinant of a matrix. We use shorthand $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$ to denote that $\boldsymbol{X}$ follows a vector elliptically contoured distribution

[2] In this chapter the vector-elliptical distribution is assumed to have a density function.

with density (3.2). The vector elliptically contoured distribution has been extensively studied in [Dav77, AF90, And92].

An example of a vector-elliptical distributed is the centered matrix normal distribution, for which the rows $\{\boldsymbol{x}^{(i)}\}_{i=1}^n \subset \mathbb{R}^p$ are i.i.d. samples from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. In this special case, $\boldsymbol{X} \sim \mathcal{VE}(\mathbf{0}, \boldsymbol{\Sigma}, \theta)$ and the density of $\boldsymbol{X}$ is given by (3.2) with $\theta(w) = (2\pi)^{-\frac{np}{2}} \exp(-\frac{1}{2}w)$. Specifically

$$f_{\boldsymbol{X}}(\boldsymbol{X}) = \det(\boldsymbol{\Sigma})^{-n/2}(2\pi)^{-\frac{np}{2}} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^T\right)\right). \tag{3.3}$$

Given a data matrix $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$, the sample mean is

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^{(i)} = \frac{1}{n}\boldsymbol{X}^T\mathbf{1}$$

and the sample covariance matrix $\boldsymbol{S}$ is

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}})(\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}})^T = \frac{1}{n-1}(\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^T)^T(\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^T). \tag{3.4}$$

The sample correlation matrix $\boldsymbol{R}$ is defined as:

$$\boldsymbol{R} = \text{diag}(\boldsymbol{S})^{-\frac{1}{2}}\boldsymbol{S}\text{diag}(\boldsymbol{S})^{-\frac{1}{2}}, \tag{3.5}$$

where $\text{diag}(\boldsymbol{A})$ for a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the diagonal part of $\boldsymbol{A}$ and $\boldsymbol{B}^{-1/2}$ for a diagonal matrix $\boldsymbol{B}$ is a diagonal matrix by raise every diagonal element of $\boldsymbol{B}$ to the power $-1/2$. Since $\boldsymbol{R}$ is not necessarily invertible, we define $\boldsymbol{R}^\dagger$ as the Moore-Penrose pseudo-inverse of $\boldsymbol{R}$ and define the sample partial correlation matrix $\boldsymbol{P}$ by

$$\boldsymbol{P} = \text{diag}(\boldsymbol{R}^\dagger)^{-\frac{1}{2}}\boldsymbol{R}^\dagger\text{diag}(\boldsymbol{R}^\dagger)^{-\frac{1}{2}}. \tag{3.6}$$

For convenience we define $\{\boldsymbol{\Psi}^{(k)}\}_{k\in\{\boldsymbol{R},\boldsymbol{P}\}}$ to be matrices such that $\boldsymbol{\Psi}^{(\boldsymbol{R})} = \boldsymbol{R}$ and $\boldsymbol{\Psi}^{(\boldsymbol{P})} = \boldsymbol{P}$. Given a threshold $\rho \in [0, 1)$ define the undirected graph induced by thresholding $\boldsymbol{\Psi}^{(k)}$, denoted by $\mathcal{G}_\rho(\boldsymbol{\Psi}^{(k)})$, as follows. The vertex set of graph $\mathcal{G}_\rho(\boldsymbol{\Psi}^{(k)})$ is $\mathcal{V}^{(k)} = [p] := \{1, 2, \cdots, p\}$ and the edge set is $\mathcal{E}^{(k)} \subset \mathcal{V}^{(k)} \times \mathcal{V}^{(k)}$, where there is an edge between i and j $(i \neq j)$, i.e., $(i, j) \in \mathcal{E}^{(k)}$, if $|\Psi_{ij}^{(k)}| \geq \rho$. Let $\boldsymbol{\Phi}^{(k)}(\rho)$ be the adjacency matrices associated with the graph $\mathcal{G}_\rho(\boldsymbol{\Psi}^{(k)})$, defined as $\Phi_{ij}^{(k)}(\rho) = 1(|\Psi_{ij}^{(k)}| \geq \rho)$ for $i \neq j$, where $1(\cdot)$ is the indicator function. We call $\mathcal{G}_\rho(\boldsymbol{\Psi}^{(k)})$ the empirical correlation graph and the empirical partial correlation graph respectively when $\boldsymbol{\Psi}^{(k)} = \boldsymbol{R}$ and $\boldsymbol{\Psi}^{(k)} = \boldsymbol{P}$. The dependence of $\Phi_{ij}^{(k)}(\rho)$ on ρ will be suppressed if it's clear from context.

The focus of this chapter is correlation screening for which the objective is to identify connected

vertices or vertices of prescribed degree in $\mathcal{G}_\rho(\mathbf{\Psi}^{(k)})$. The number of vertices of varying degrees specifies the degree distribution. Characterization of the distributions of these counting statistics is the main contribution of the chapter. More specifically, for the graph $\mathcal{G}_\rho(\mathbf{\Psi}^{(k)})$ with $k = \boldsymbol{R}$ or $k = \boldsymbol{P}$, the degree of vertex i is defined as $\sum_{j=1,j\neq i}^{p} \Phi_{ij}^{(k)}(\rho)$. For $1 \leq \delta \leq p-1$, the total number of vertices with degree exactly δ (at least δ), denoted by $N_{\breve{V}_\delta}^{(k)}$ ($N_{V_\delta}^{(k)}$), are of particular interest in this chapter. Moreover, let Γ_δ be a star shaped graph with δ edges. For $2 \leq \delta \leq p-1$, the number of subgraphs in $\mathcal{G}_\rho(\mathbf{\Psi}^{(k)})$ that are isomorphic to Γ_δ is denoted by $N_{E_\delta}^{(k)}$. Moreover, we define $N_{E_1}^{(k)}$ to be twice of number of edges in $\mathcal{G}_\rho(\mathbf{\Psi}^{(k)})$. $N_{E_\delta}^{(k)}$ is referred as *star subgraph counts*. The following is an example to illustrate the 6 quantities defined.

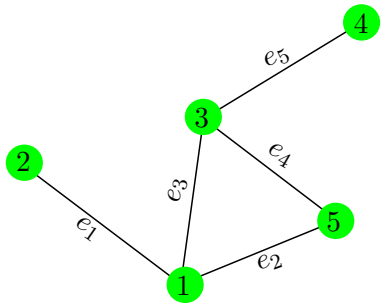


Figure 3.1: A graph with 5 vertices and 5 edges.

Example 3.2.1. Let Figure 3.1 represent an empirical partial correlation graph. For this graph the number of vertices of degree 2 is $N_{\breve{V}_2}^{(\boldsymbol{P})} = 1$ and the number of vertices of degree at least 2 is $N_{V_2}^{(\boldsymbol{P})} = 3$. The number of subgraphs isomorphic to Γ_3 is $N_{E_3}^{(\boldsymbol{P})} = 2$. The number of connected vertices is $N_{V_1}^{(\boldsymbol{P})} = 5$, and $N_{E_1}^{(\boldsymbol{P})} = 10$ as there are 5 edges.

Consider now the case where the number of sample n is fixed while there is a sequence of data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with increasing number of dimension p. Then this induces a sequence of random graphs $G_\rho(\mathbf{\Psi}^{(k)})$ with increasing number of vertices. This chapter derives finite sample compound Poisson characterization of the distribution of the 6 random quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ for sufficiently large dimension p and correlation threshold ρ, under the some sparsity assumption on the dispersion parameter $\mathbf{\Sigma}$.

3.2.2 A unified theorem

As discussed in the previous subsection, the main theorem of the chapter is to study the distribution of $\bar{N}_\delta$, where $\bar{N}_\delta$ is a generic random variable of the 6 random quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$. Indeed, our main theorem establishes $\bar{N}_\delta$ converges in distribution to compound Poisson. We will begin by defining necessary quantities and then state our main theorem at the end of this subsection.

For any positive number λ and a probability distribution $\boldsymbol{\zeta}$ supported on positive integers, let $\mathrm{CP}(\lambda, \boldsymbol{\zeta})$ be the corresponding compound Poisson distribution, i.e. $\mathrm{CP}(\lambda, \boldsymbol{\zeta})$ is the distribution of $Z = \sum_{i=1}^{N} Z_i$, where N is distributed as a Poisson random variable with mean λ, $Z_i \overset{i.i.d.}{\sim} \boldsymbol{\zeta}$ and N is independent of each Z_i. Here the random variable N measures the number of occurrences of increments and the distribution $\boldsymbol{\zeta}$ characterize the distribution of each increment.

The parameters of the limiting compound Poisson distribution in our main theorem involves random geometric graph. Given a set of points $\{\boldsymbol{v}_i\}_{i=1}^{\delta}$ in $\mathbb{R}^{n-2}$, denote by $\mathbf{Ge}\left(\{\boldsymbol{v}_i\}_{i=1}^{\delta}, r; \delta, n-2\right)$ the geometric graph with radius r, defined as follows. The vertex set of the graph is $\{\boldsymbol{v}_i\}_{i=1}^{\delta}$, and there is an edge between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ if $\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2 \le r$. Denote by $\mathbf{NMD}\left(\{\boldsymbol{v}_i\}_{i=1}^{\delta}, r; \delta, n-2\right)$ the number of vertices of maximum degree $\delta - 1$ in $\mathbf{Ge}\left(\{\boldsymbol{v}_i\}_{i=1}^{\delta}, r; \delta, n-2\right)$. Let $\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}$ are i.i.d. $\mathrm{unif}(B^{n-2})$, the uniform distribution in B^{n-2}, which denotes unit ball in $\mathbb{R}^{n-2}$. For the random geometric graph $\mathbf{Ge}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2\right)$, denote the probability that there are exactly $\ell - 1$ vertices of maximum degree $\delta - 1$ by

$$\alpha_\ell := \mathbb{P}\left(\mathbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2\right) = \ell - 1\right), \quad \forall \ell \in [\delta + 1], \tag{3.7}$$

and define a probability measure $\boldsymbol{\zeta}_{n,\delta}$ on $[\delta + 1]$:

$$\boldsymbol{\zeta}_{n,\delta}(\ell) := (\alpha_\ell/\ell) \Big/ \left(\sum_{\ell=1}^{\delta+1} (\alpha_\ell/\ell)\right), \quad \forall \ell \in [\delta + 1]. \tag{3.8}$$

This $\boldsymbol{\zeta}_{n,\delta}$ is the distribution of each increment for the limiting compound Poisson distribution.

To ensure convergence in our main theorem, some sparsity conditions on the dispersion parameter $\boldsymbol{\Sigma}$ are imposed. A matrix is row-κ sparse if every row of it has at most κ nonzero elements. The next is a stronger sparsity than row-κ sparse.

Definition 3.2.2 ((τ, κ) sparsity)**.** A p by p dimensional symmetric matrix is call (τ, κ) sparse and it's row-κ sparse and its right-bottom $p - \tau$ by $p - \tau$ sub-matrix is diagonal.

Another relevant quantity is the normalized determinant.

Definition 3.2.3 (Normalized determinant)**.** For any symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, its normalized determinant $\mu(\boldsymbol{A})$ is defined by

$$\mu(\boldsymbol{A}) := \prod_{i=1}^{p} \frac{\lambda_i(\boldsymbol{A})}{\lambda_p(\boldsymbol{A})} = \frac{\det(\boldsymbol{A})}{(\lambda_{\max}(\boldsymbol{A}))^p},$$

where $\lambda_1(\boldsymbol{A}) \le \lambda_2(\boldsymbol{A}) \le \cdots \le \lambda_p(\boldsymbol{A})$ are the eigenvalues of $\boldsymbol{A}$.

Denote by $\boldsymbol{A}_{\mathcal{I}}$ for $\mathcal{I} \subset [p]$ to be the set of all $|\mathcal{I}|$ by $|\mathcal{I}|$ submatrix of $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ by extracting corresponding rows and columns indexed by $\mathcal{I}$. $\boldsymbol{A}_{\mathcal{I}}$ consist of $|\mathcal{I}|!$ matrices and they

are all equivalent to each other up to a permutation applying simultaneously to both rows and columns. Define the *local normalized determinant of degree* m of a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ to be $\mu_m(\boldsymbol{A}) = \min\{\mu(\boldsymbol{A}_{\mathcal{I}}) : \mathcal{I} \subset [p], |\mathcal{I}| = m\}$, where $\mu(\boldsymbol{A}_{\mathcal{I}})$ is well defined since $\mu(\cdot)$ is invariant to simultaneously applying a permutation to both rows and columns of its argument. For $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ further define

$$\mu_{n,m}(\boldsymbol{A}) := \begin{cases} [\mu_m(\boldsymbol{A})]^{-\frac{n-1}{2}}, & \boldsymbol{A} \text{ symmetric positive definite but not diagonal,} \\ 1, & \boldsymbol{A} \text{ symmetric positive definite and diagonal.} \end{cases} \tag{3.9}$$

By definition $\mu_{n,m}(\boldsymbol{A}) \in [1, \infty)$.

We are now in a good position to state our main theorem, which states when $p \to \infty$, if the threshold ρ is chosen to approach 1 at a particular rate, then the sequence of each of the 6 quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$, converges to a compound Poisson random variable in distribution.

Theorem 3.2.4 (Compound Poisson Limit)**.** *Let $n \geq 4$ and δ be fixed positive integers. Let $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Choose threshold ρ as a function of p such that $c_n 2^{\frac{n}{2}} p^{1+\frac{1}{\delta}} (1-\rho)^{\frac{n-2}{2}} \to e_{n,\delta}$ as $p \to \infty$, where $c_n = \frac{\Gamma((n-1)/2)}{(n-2)\sqrt{\pi}\Gamma((n-2)/2)}$ and $e_{n,\delta}$ is some positive constant that possibly depends on n and δ. Denote $\lambda_{n,\delta}(e_{n,\delta}) = \frac{1}{\delta!}(e_{n,\delta})^\delta \sum_{\ell=1}^{\delta+1} \frac{\alpha_\ell}{\ell}$. Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ_p, κ_p) sparse with $\lim_{p\to\infty} \frac{\tau_p}{p} + \mu_{n,2\delta+2}(\boldsymbol{\Sigma}) \frac{\kappa_p}{p} \to 0$. Then with $\bar{N}_\delta$ a generic random variable in the set $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$:*

$$\bar{N}_\delta \to CP(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta}) \text{ in distribution as } p \to \infty. \tag{3.10}$$

Remark 3.2.5 (Relaxed sparsity assumption in empirical correlation graph)**.** If only random quantities in the empirical correlation graph are of concern, then the (τ_p, κ_p) sparsity assumption can be relaxed to row-κ sparsity. Specifically, the last two sentences in Theorem 3.2.4 can be replaced by the following.

Suppose $\boldsymbol{\Sigma}$ is row-κ_p sparse with $\lim_{p\to\infty} \mu_{n,2\delta+2}(\boldsymbol{\Sigma}) \frac{\kappa_p}{p} \to 0$. Then with $\tilde{N}_\delta$ a generic random variable in the set $\{N_i^{(k)} : k = \boldsymbol{R}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$:

$$\tilde{N}_\delta \to CP(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta}) \text{ in distribution as } p \to \infty. \tag{3.11}$$

□

Remark 3.2.6. The condition $c_n 2^{\frac{n}{2}} p^{1+\frac{1}{\delta}} (1-\rho)^{\frac{n-2}{2}} \to e_{n,\delta} > 0$ is equivalent to

$$p^{\frac{2}{n-2}\left(1+\frac{1}{\delta}\right)}(1-\rho) \to \left(\frac{e_{n,\delta}}{c_n 2^{\frac{n}{2}}}\right)^{\frac{2}{n-2}} = \frac{1}{2}\left(\frac{e_{n,\delta}}{2c_n}\right)^{\frac{2}{n-2}},$$

which indicates that the rate $\rho \to 1$ is $p^{-\frac{2}{n-2}\left(1+\frac{1}{\delta}\right)}$. One possible choice for such a threshold is by replacing the convergence sign in the preceding display with equal sign, which yields

$$\rho = 1 - \frac{1}{2}\left(\frac{e_{n,\delta}}{2c_n p^{1+\frac{1}{\delta}}}\right)^{\frac{2}{n-2}}. \tag{3.12}$$

□

The proofs of Theorem 3.2.4 and Remark 3.2.5 will be presented in Subsection 3.3. Since for discrete random variables convergence in distribution is equivalent to convergence in total variation, (3.11) is equivalent to:

$$d_{TV}\left(\mathscr{L}\left(\bar{N}_\delta\right), CP(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta})\right) \to 0 \text{ as } p \to \infty,$$

where $\mathscr{L}(\cdot)$ represents the probability distribution of the random variable in its argument, and $d_{TV}(\cdot,\cdot)$ is the total variation distance between two probability distributions. A quantitative version of Theorem 3.2.4 establishing upper bound on the total variation distance between $\mathscr{L}\left(\bar{N}_\delta\right)$ and a compound Poisson for finite p will be presented in Theorem 3.3.11.

In Theorem 3.2.4 assumptions on (τ,κ) sparsity and the quantity $\mu_{n,m}(\boldsymbol{\Sigma})$ are imposed. In the next two subsections, we will present examples of these two new definitions. We also compare each of the new definitions with relevant classical quantities to illustrate that the assumptions on them are not restrictive.

3.2.3 (τ,κ) sparsity

In the current subsection examples of (τ,κ) sparse matrix and the comparison between (τ,κ) sparsity and other sparsity are presented to elaborate the (τ,κ) sparsity imposed in the Theorem 3.2.4.

The matrix below in (3.13) is an example of (τ,κ) sparse matrix with $\tau = 2, \kappa = 3$. This 5×5 symmetric matrix is $(2,3)$ sparse since each of the first 2 rows has at most 3 nonzero elements and

the right-bottom 3×3 sub-matrix is diagonal.

$$\begin{pmatrix} 5 & 0 & 2 & 0 & 1 \\ 0 & 3 & 2017 & 0 & 0 \\ 2 & 2017 & 6 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 \\ 1 & 0 & 0 & 0 & 8 \end{pmatrix} \tag{3.13}$$

If the adjacency matrix of a graph $(\mathcal{V}, \mathcal{E})$ is (τ, κ) sparse, then the vertices $\mathcal{V}$ can be partitioned into two disjoint subsets $\mathcal{V}_1$ and $\mathcal{V}_2$ with the following properties: 1) $|\mathcal{V}_1| \leq \tau$; 2) there is no edge between any two vertices in $\mathcal{V}_2$; 3) the degree of any vertex in $\mathcal{V}_1$ is no more than $\kappa - 1$; 4) edges connecting vertex in $\mathcal{V}_1$ and $\mathcal{V}_2$ are allowed (but not necessarily existent).

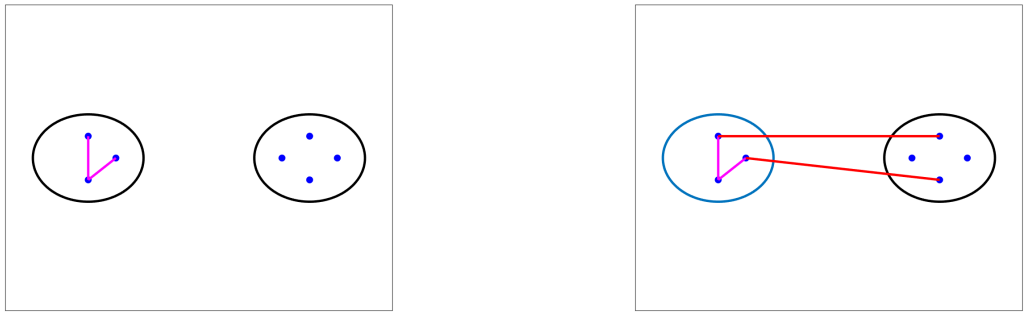
When the dispersion parameter $\mathbf{\Sigma}$ is row-κ sparse, [HR11] studied the mean of $N_{E_1}^{(\boldsymbol{R})}$ and $N_{V_1}^{(\boldsymbol{R})}$ and obtained limits of the probability when they are respectively nonzero. [HR12] extends these results to empirical partial correlation graph when the dispersion parameter $\mathbf{\Sigma}$ is assumed to be, up to a row-column permutations, block-τ sparse, i.e., there exists a permutation matrix $\boldsymbol{T}$ such that

$$\boldsymbol{T}\mathbf{\Sigma}\boldsymbol{T}^T = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \boldsymbol{D}_{p-\tau} \end{pmatrix} \tag{3.14}$$

where $\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}^T = \mathbf{0} \in \mathbb{R}^{\tau \times (p-\tau)}$ and $\boldsymbol{D}_{p-\tau} \in \mathbb{R}^{(p-\tau)\times(p-\tau)}$ is some diagonal matrix. In Theorem 3.2.4 $\mathbf{\Sigma}$ is imposed to be (τ, κ) sparse after some row-column permutation, i.e. there exists a permutation matrix $\boldsymbol{T}$ such that (3.14) holds with $\boldsymbol{D}_{p-\tau} \in \mathbb{R}^{(p-\tau)\times(p-\tau)}$ some diagonal matrix and with the first τ rows $(\mathbf{\Sigma}_{11}\ \mathbf{\Sigma}_{12})$ being row-κ sparse. It should be clear that (τ, κ) sparsity is more general than the block sparsity since there is no restriction on $\mathbf{\Sigma}_{12} = 0$. We make this comparison precise in the next paragraph.

Obviously, (τ, κ) sparsity reduces to block-τ sparsity in [HR12] as a special case. Indeed, every block-τ sparse matrix is (τ, κ) sparse with $\kappa = \tau$. (τ, κ) sparsity with $\kappa = \tau$, nevertheless, allows non-zeros in the top-right submatrix, which mean more possible correlations between the variables than block-τ sparsity in correlation graphical models. To see this, consider the associated graphical model $\mathcal{G}_0(\mathbf{\Sigma})$.[3] In Figure 3.2, nodes represent the variables and edges represent the correlation between variables. The left panel is a graphical model associated to the block-3 sparse assumption, while the right panel satisfies (τ, κ) sparsity with $(\tau, \kappa) = (3, 3)$. The later has more correlations (the red edges) across the two sets of variables in the 2 circles. (τ, κ) sparsity with $\kappa > \tau$ will further enrich the possible correlations between variables.

[3] In Section 3.2.1 we define $\mathcal{G}_\rho(\cdot)$ as the induced graph with thresholding by ρ for a matrix. Here $\mathcal{G}_0(\cdot)$, is the induced graph for the matrix without thresholding.



(a) block sparsity (b) (τ, κ) sparsity

Figure 3.2: Diagram of the correlation graph $\mathcal{G}_0(\boldsymbol{\Sigma})$ for $p = 7$ dimensional distributions with two different 7×7 covariance matrices. The left panel is associated with a block-3 sparse assumption on $\boldsymbol{\Sigma}$. Only the $\tau = 3$ variables in the group inside the left circle are correlated: there is no correlation (edge) between the remaining 4 variables in the right circle and there is no correlation across the two sets of variables in different circles. The right panel is associated with $(\tau, \kappa) = (3, 3)$ sparsity on $\boldsymbol{\Sigma}$, where two additional edges, representing correlations between variables, exist across the two groups.

On the other hand, (τ, κ) sparsity is a stronger assumption than row-κ sparsity, since every (τ, κ) sparse matrix is row-κ sparse. (τ, κ) sparsity is thus an intermediate level of sparsity lying between block sparsity and row sparsity.

For a unified framework in empirical correlation and partial correlation graph, we supposed that the dispersion parameter $\boldsymbol{\Sigma}$, after some row-column permutations, is (τ, κ) sparse as stated in Theorem 3.2.4.

Remark 3.2.7. Recall that we are interested in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$, quantities that are invariant under permutation of the p-dimensional variables. Since permutation of the variables is equivalent to the row-column permutation to the dispersion parameter $\boldsymbol{\Sigma}$, without loss of generality, we can assume the variables are permuted such that $\boldsymbol{\Sigma}$ is (τ, κ) sparse. □

3.2.4 Local normalized determinant

In this subsection, we provide further details on the normalized determinant and local normalized determinant defined in Subsection 3.2.2. Sufficient conditions to control the quantity $\mu_{n,m}(\cdot)$ used in Theorem 3.2.4 in terms of eigenvalues or condition numbers are also discussed.

Observe applying the same permutation simultaneously to both rows and columns of a matrix does not change its normalized determinant. It is also obvious $\mu(\boldsymbol{A}) \in (0, 1]$ and, $\mu(\boldsymbol{A}) = 1$ if and only if $\boldsymbol{A}$ is a multiple of $\boldsymbol{I}_p$. Moreover, $\mu(\boldsymbol{A})$ is close to 1 and hence bounded away from 0, as long as all eigenvalues concentrate around a positive number. Below is an example of a sequence of

symmetric positive definite matrices with well-concentrated eigenvalues such that their normalized determinants are uniformly bounded away from 0.

Example 3.2.8. Let $\{\alpha_i\}_{i=1}^p$ be positive. Let $\{\beta_i\}_{i=1}^\infty$ be a positive, decreasing sequences such that $\sum_{i=1}^{\infty} \beta_i < \infty$. Consider $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ be a symmetric positive definite matrix with eigenvalues $\lambda_i = \alpha_p \exp(-\beta_i)$ for $1 \le i \le p-1$ and $\lambda_p = \alpha_p$. Then

$$\mu(\boldsymbol{A}) = \exp\left(-\sum_{i=1}^{p-1} \beta_i\right).$$

Consider now p is increasing, i.e. consider a sequence of matrices $\boldsymbol{A}$ of the above properties with increasing dimension. For this sequence of matrices $\mu(\boldsymbol{A}) \ge \exp\left(-\sum_{i=1}^{\infty} \beta_i\right) > 0$, i.e. $\mu(\boldsymbol{A})$ is bounded uniformly away from 0.

For local normalized determinant, it follows immediately by interlacing property (cf. Theorem 8.1.7 in [GVL12]) that $\mu_m(\boldsymbol{A})$ is decreasing with respect to $m \in [p]$ for any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$. Thus $\mu_{n,m}(\boldsymbol{A})$ defined in (3.9) is increasing with respect to $m \in [p]$.

It turns out the local normalized determinant of the dispersion matrix $\boldsymbol{\Sigma}$ will play an important role in our study of the distribution of the six quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$. Heuristically, when $\delta \ge 2$, $N_{E_\delta}^{(k)}$, as a sum of indicator functions of whether a subgraph of $\delta+1$ vertices is isomorphic Γ_δ, has the local property in the sense that each term in the summation involves only $\delta+1$ variables, and thus each pair of two such terms involves at most $2(\delta+1)$ variables. So heuristically $\mu_{2(\delta+1)}(\boldsymbol{\Sigma})$ controls the correlation between two indicator terms in the summation of $N_{E_\delta}^{(k)}$, which has impact on the convergence of $N_{E_\delta}^{(k)}$ to a compound Poisson. $N_{\breve{V}_\delta}^{(k)}$ and $N_{V_\delta}^{(k)}$ will be shown to have similar local property as $N_{E_\delta}^{(k)}$ in Lemma 3.3.8.

It is imposed in Theorem 3.2.4 that $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})\frac{\kappa_p}{p} \to 0$, which holds when $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})$ is either bounded or increases in a rate $o(\frac{p}{\kappa_p})$. We provide some sufficient condition on well-studied concepts like condition number or eigenvalues of $\boldsymbol{\Sigma}$ to control $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})$ in Section 3.7. Specifically, we show $\mu_{n,m}(\cdot)$ is upper bounded by powers of the condition number of its argument, and provide two sets of sufficient conditions to guarantee uniform boundedness of $\mu_{n,m}(\cdot)$ for a sequence of symmetric positive definite matrices.

3.3 Non-asymptotic compound Poisson approximation

In this section we establish a non-asymptotic compound Poisson approximation for $\mathscr{L}(\bar{N}_\delta)$, with $\bar{N}_\delta$ a generic random variable in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$, based on which we present the proof of Theorem 3.2.4. In Subsection 3.3.1 scores representations for our model are

introduced and these scores are key concepts in the following development. Subsection 3.3.2 is devoted to provide an equivalent formulation of our model in terms of random geometric graphs, where parameters of the non-asymptotic compound Poisson approximation are defined. With the preparation of the first two subsections, Subsection 3.3.3 presents the non-asymptotic compound Poisson approximation for star subgraph counts $N_{E_\delta}^{(\boldsymbol{R})}$ and in Subsection 3.3.4 it is developed that all 6 quantities in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ are close in L^1 distance. Then combining results in Subsection 3.3.3 and Subsection 3.3.4, a theorem on non-asymptotic compound Poisson approximation is obtained, which then implies Theorem 3.2.4 in Subsection 3.3.5.

3.3.1 Score representations of sample correlation and partial correlation

In this subsection, the $\boldsymbol{U}$-score for the empirical correlation graph and the $\boldsymbol{Y}$-score for the empirical partial correlation graph are defined. These scores will serve as the vertices set on which the random geometric graphs are constructed in Subsection 3.3.2.

We first present two useful reductions to our model. The first reduction is to reduce $\mathscr{L}(\boldsymbol{X})$ from vector elliptically contoured distribution as in (3.2) to a centered matrix normal distribution as in (3.3).

Reduction to centered matrix normal data matrix

The following lemma is an immediate result of the theorem in Section 6 of [And92].

Lemma 3.3.1. *Suppose $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Then the distribution of $\boldsymbol{R}$ defined in* (3.5) *is invariant to θ and $\boldsymbol{\mu}$.*

Since the quantities of interest in this chapter are $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$, functions of $\boldsymbol{R}$, their distribution are also invariant to θ and $\boldsymbol{\mu}$ when $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Specifically, we may choose $\theta(w) = (2\pi)^{-\frac{np}{2}} \exp(-\frac{1}{2}w)$ and $\boldsymbol{\mu} = \mathbf{0}$ such that $\boldsymbol{X}$ has density (3.3). That is, $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$, the rows of $\boldsymbol{X}$, can be taken as i.i.d. samples from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, without changing the distribution of any of the quantity $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$. Thus for the rest of the chapter we will suppose $\boldsymbol{X}$ has density (3.3) and refer to the dispersion parameter $\boldsymbol{\Sigma}$ as the population/theorectical covariance matrix. As a consequence, $\Sigma_{ij} = 0$ implies variable i and variable j are independent. In particular, the different sparsity conditions discussed in Subsection 3.2.3 imposed on $\boldsymbol{\Sigma}$ induces different independence structures between variables in the reduced model.

The second reduction represents the sample covariance matrix defined in (3.4) by a sample second moment of the projected data.

Reduction from sample covariance matrix to sample second moment

It is shown in Theorem 3.3.2 [And03] that the sample covariance matrix of n i.i.d. normal distribution is identical to the sample second moment of $n-1$ i.i.d. zero-mean normal random vectors. Specifically, define the orthogonal $n \times n$ matrix $\boldsymbol{H} = [n^{-\frac{1}{2}}\mathbf{1}, \boldsymbol{H}_{2:n}]^T$. The matrix $\boldsymbol{H}_{2:n}$ can be obtained by Gramm-Schmidt orthogonalization and satisfies the properties

$$\mathbf{1}^T \boldsymbol{H}_{2:n} = \mathbf{0}, \quad \boldsymbol{H}_{2:n}^T \boldsymbol{H}_{2:n} = \boldsymbol{I}_{n-1}.$$

Then

$$\tilde{\boldsymbol{X}} = \boldsymbol{H}_{2:n}\boldsymbol{X} \in \mathbb{R}^{(n-1)\times p} \tag{3.15}$$

have i.i.d. rows $\{\tilde{\boldsymbol{x}}^{(i)}\}_{i=1}^{n-1}$ drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Moreover the sample covariance matrix defined in (3.4)

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n-1} \tilde{\boldsymbol{x}}^{(i)}(\tilde{\boldsymbol{x}}^{(i)})^T = \frac{1}{n-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}. \tag{3.16}$$

That is, the sample covariance matrix of the data matrix $\boldsymbol{X}$, defined in (3.4), is the same as the sample second moment of the data matrix $\tilde{\boldsymbol{X}}$.

We are now in a good position to define and analyze the scores representation of $\boldsymbol{R}$ and $\boldsymbol{P}$. From the second reduction above, $\{\tilde{\boldsymbol{x}}^{(i)}\}_{i=1}^{n-1} \subset \mathbb{R}^p$, the rows of $\tilde{\boldsymbol{X}}$, are i.i.d. copy from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Let $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^p$ be the columns of $\tilde{\boldsymbol{X}}$. Then $\boldsymbol{u}_i := \frac{\tilde{\boldsymbol{x}}_i}{\|\tilde{\boldsymbol{x}}_i\|_2} \in \mathbb{R}^{n-1}$ has distribution unif(S^{n-2}) for $i \in [p]$. Denote

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p] \in \mathbb{R}^{(n-1)\times p}$$

and it follows from (3.16) and (3.5) that

$$\boldsymbol{R} = \boldsymbol{U}^T\boldsymbol{U}. \tag{3.17}$$

The $\boldsymbol{U}$ in equation (3.17) is referred to as $\boldsymbol{U}$-score of the sample correlation matrix [HR11, HR12]. Note our formulation of $\boldsymbol{U}$-score is slightly different from the formulation in [HR11, HR12], but it's an easy task to check the two different formulations are equivalent. Moreover, as discussed in the first reduction above, $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are independent provided the population covariance $\Sigma_{ij} = 0$.

The normalized outer product of $\boldsymbol{U}$, defined by

$$\boldsymbol{B} = \frac{n-1}{p}\boldsymbol{U}\boldsymbol{U}^T \in \mathbb{R}^{(n-1)\times(n-1)} \tag{3.18}$$

will play an important role in the analysis of empirical partial correlation graph.

Lemma 3.3.2. *Let $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$ and $p \geq n$. Then $\boldsymbol{B}$ is invertible with probability* 1.

By Lemma 1 in [HR12], provided $\boldsymbol{U}\boldsymbol{U}^T$ is invertible or, equivalently, $\boldsymbol{B}$ is invertible,

$$\boldsymbol{R}^\dagger = \boldsymbol{U}^T[\boldsymbol{U}\boldsymbol{U}^T]^{-2}\boldsymbol{U} = \left(\frac{p}{n-1}\right)^2 \boldsymbol{U}^T\boldsymbol{B}^{-2}\boldsymbol{U}. \tag{3.19}$$

It follows from Lemma 3.3.2 that equation (3.19) holds $a.s.$. Define $\boldsymbol{A} = \boldsymbol{B}^{-1}$, $\bar{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{U}$ and hence

$$\boldsymbol{R}^\dagger = \left(\frac{p}{n-1}\right)^2 \bar{\boldsymbol{Y}}^T\bar{\boldsymbol{Y}} \quad a.s.. \tag{3.20}$$

Further define

$$\begin{aligned}
&\boldsymbol{y}_i = \bar{\boldsymbol{y}}_i/\|\bar{\boldsymbol{y}}_i\|_2, \quad \forall\, i \in p,\\
&\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p] \in \mathbb{R}^{(n-1)\times p},
\end{aligned}$$

and thus

$$\boldsymbol{P} = \boldsymbol{Y}^T\boldsymbol{Y} \quad a.s. \tag{3.21}$$

by equation (3.6) and (3.20). $\boldsymbol{Y}$ in equation (3.21) is referred to as the $\boldsymbol{Y}$-score representations for sample partial correlation matrix [HR12]. Similar to $\boldsymbol{U}$-score, one can verify easily our formulation of $\boldsymbol{Y}$-score is equivalent to that in [HR12].

Let σ^{n-2} be the spherical measure on S^{n-2}, i.e. σ^{n-2} is the probability measure of uniform distribution on S^{n-2}. Denote by $f_{\boldsymbol{u}_{j_1},\boldsymbol{u}_{j_2},\cdots,\boldsymbol{u}_{j_m}}$ the joint density of j_1-th, j_2-th, …, j_m-th column of $\boldsymbol{U}$ with respect to the product measure $\otimes^m\sigma^{n-2} := \underbrace{\sigma^{n-2}\otimes\sigma^{n-2}\otimes\cdots\otimes\sigma^{n-2}}_{m}$. The next lemma establishes $f_{\boldsymbol{u}_{j_1},\boldsymbol{u}_{j_2},\cdots,\boldsymbol{u}_{j_m}}$ is bounded by $\mu_{n,m}(\boldsymbol{\Sigma})$. This highlights the role of $\mu_{n,m}(\boldsymbol{\Sigma})$ since the distribution of the six quantities $N^{(k)}_{\check{V}_\delta}$,$N^{(k)}_{V_\delta}$, $N^{(k)}_{E_\delta}$ with $k \in \{\boldsymbol{R}, \boldsymbol{P}\}$ have local property as discussed in Subsection 3.2.4, depending only on joint density of $f_{\boldsymbol{u}_{j_1},\boldsymbol{u}_{j_2},\cdots,\boldsymbol{u}_{j_m}}$ for every possible collection of $\{j_i\}_{i=1}^m$ and some constant $m = 2\delta + 2$.

Lemma 3.3.3. *Let $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Consider $\{j_i\}_{i=1}^m \subset [p]$ are distinct indexes and $\mathcal{J} = \{j_i : 1 \le i \le m\}$.*

(a)

$$\mu_{n,m}(\boldsymbol{\Sigma}_{\mathcal{J}}) \le \mu_{n,m}(\boldsymbol{\Sigma})$$

(b) The joint density of any subset of m columns of $\boldsymbol{U}$-score w.r.t. $\otimes^m\sigma^{n-1}$ is upper bounded by $\mu_{n,m}(\boldsymbol{\Sigma})$:

$$f_{\boldsymbol{u}_{j_1},\boldsymbol{u}_{j_2},\cdots,\boldsymbol{u}_{j_m}}(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m) \le \mu_{n,m}(\boldsymbol{\Sigma}_{\mathcal{J}}) \le \mu_{n,m}(\boldsymbol{\Sigma}), \quad \forall\, \boldsymbol{v}_i \in S^{n-2}, \forall\, i \in [m].$$

Here $\mu_{n,m}(\boldsymbol{\Sigma}_{\mathcal{J}})$ is well defined since $\mu_{n,m}(\cdot)$ is invariant to simultaneously applying a permutation to both rows and columns.

(c) Let $h : (S^{n-2})^m \to \mathbb{R}$ be a nonnegative Borel measurable function. Then

$$\mathbb{E}h(\boldsymbol{u}_{j_1}, \boldsymbol{u}_{j_2}, \cdots, \boldsymbol{u}_{j_m}) \le \mu_{n,m}(\boldsymbol{\Sigma}_{\mathcal{J}})\mathbb{E}h(\boldsymbol{u}'_{j_1}, \boldsymbol{u}'_{j_2}, \cdots, \boldsymbol{u}'_{j_m}) \le \mu_{n,m}(\boldsymbol{\Sigma})\mathbb{E}h(\boldsymbol{u}'_{j_1}, \boldsymbol{u}'_{j_2}, \cdots, \boldsymbol{u}'_{j_m}),$$

where $\{\boldsymbol{u}'_{j_\ell}\}_{\ell=1}^m$ are i.i.d. distributed as unif(S^{n-2}).

Remark 3.3.4. In Lemma 3.3.3 (a), since $\boldsymbol{\Sigma}_{\mathcal{J}}$ is m by m symmetric positive definite matrix, by the definition in (3.9),

$$\mu_{n,m}(\boldsymbol{\Sigma}_{\mathcal{J}}) = \begin{cases} [\mu(\boldsymbol{\Sigma}_{\mathcal{J}})]^{-\frac{n-1}{2}}, & \boldsymbol{\Sigma}_{\mathcal{J}} \text{ not diagonal,} \\ 1, & \boldsymbol{\Sigma}_{\mathcal{J}} \text{ diagonal.} \end{cases}$$

□

The proof of Lemma 3.3.3 (b) is deferred to Appendix 3.8.2. (c) immediately follows from (b) by writing expectation as integrals. (a) follows trivially from $\mu_m(\boldsymbol{\Sigma}_{\mathcal{J}}) = \mu(\boldsymbol{\Sigma}_{\mathcal{J}}) \ge \mu_m(\boldsymbol{\Sigma})$.

Lemma 3.3.3 (c) is useful since when calculating expectation of nonnegative function of any m columns of $\boldsymbol{U}$, one may always assume the associated columns $\{\boldsymbol{u}_j\}$ are independent unif(S^{n-1}) with the cost of an additional multiplicative factor $\mu_{n,m}(\boldsymbol{\Sigma})$.

3.3.2 Random pseudo geometric graph

In this subsection we define random pseudo geometric graph and provide an equivalent formulation of our model, of which the vertices set are the scores defined in Subsection 3.3.1. We also define the increment distribution of the compound Poisson that non-asymptotically approximates 6 random quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$.

From equation (3.17) and the fact that columns of $\boldsymbol{U}$ have Euclidean norm 1,

$$R_{ij} = \boldsymbol{u}_i^T \boldsymbol{u}_j = 1 - \frac{\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2^2}{2} = \frac{\|\boldsymbol{u}_i + \boldsymbol{u}_j\|_2^2}{2} - 1. \tag{3.22}$$

For a threshold $\rho \in [0, 1)$, define $r_\rho := \sqrt{2(1-\rho)} \in (0, \sqrt{2}]$. By equation (3.22),

$$\{R_{ij} \ge \rho\} = \{\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 \le \sqrt{2(1-\rho)}\} = \{\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 \le r_\rho\}.$$

and similarly,

$$\{R_{ij} \le -\rho\} = \{\|\boldsymbol{u}_i + \boldsymbol{u}_j\|_2 \le r_\rho\}.$$

The two preceding displays yield

$$\{|R_{ij}| \geq \rho\} = \{\|\boldsymbol{u}_i + \boldsymbol{u}_j\|_2 \leq r_\rho\} \cup \{\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 \leq r_\rho\}. \tag{3.23}$$

An entirely analogous derivation shows for empirical correlation graph,

$$\{|P_{ij}| \geq \rho\} = \{\|\boldsymbol{y}_i + \boldsymbol{y}_j\|_2 \leq r_\rho\} \cup \{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 \leq r_\rho\}. \tag{3.24}$$

Based on (3.23) and (3.24), presented in [HR11, HR12], we now introduce some novel geometric contents connecting to random geometric graph.

Note (3.23) indicates, $\{|R_{ij}| \geq \rho\}$, the event when sample correlation between i-th and j-th variables exceed the threshold ρ, or equivalently, the event when there exists a edge connecting the i-th and j-th vertices in the empirical correlation graphs $\mathcal{G}_\rho(\boldsymbol{\Psi}^{(\boldsymbol{R})})$, is the same as the event that $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$, the associated $\boldsymbol{U}$-score, lie in some geometric set on $S^{n-2} \times S^{n-2}$. This insight provide an equivalent way to construct $\mathcal{G}_\rho(\boldsymbol{\Psi}^{(\boldsymbol{R})})$ through the $\boldsymbol{U}$-scores. Similar interpretation can be drawn for (3.24). Such equivalent construction is made formal in the next few paragraphs.

Definition 3.3.5 (Pseudo geometric graph)**.** Given $m \geq 2$ and a set of points $\{\boldsymbol{v}_i\}_{i=1}^m$ in $\mathbb{R}^{\mathscr{N}}$, denote by $\mathbf{PGe}\left(\{\boldsymbol{v}_i\}_{i=1}^m, r; m, \mathscr{N}\right)$ the pseudo geometric graph with radius r, defined as follows. The vertex set of the graph is $\{\boldsymbol{v}_i\}_{i=1}^m$, and there is an edge between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ if $\mathbf{dist}(\boldsymbol{v}_i, \boldsymbol{v}_j) := \min\{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2, \|\boldsymbol{v}_i + \boldsymbol{v}_j\|_2\} \leq r$.

It's easy to verify that $\mathbf{dist}(\cdot, \cdot)$ has the following properties: for $\forall \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3 \in \mathbb{R}^m$,

1. $\mathbf{dist}(\boldsymbol{v}_1, \boldsymbol{v}_2) \geq 0$;
2. $\mathbf{dist}(\boldsymbol{v}_1, \boldsymbol{v}_2) = 0$ if only if $\boldsymbol{v}_1 = \boldsymbol{v}_2$ or $\boldsymbol{v}_1 = -\boldsymbol{v}_2$;
3. $\mathbf{dist}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \mathbf{dist}(\boldsymbol{v}_2, \boldsymbol{v}_1)$ and $\mathbf{dist}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \mathbf{dist}(\boldsymbol{v}_1, -\boldsymbol{v}_2)$
4. $\mathbf{dist}(\boldsymbol{v}_1, \boldsymbol{v}_2) \leq \mathbf{dist}(\boldsymbol{v}_1, \boldsymbol{v}_3) + \mathbf{dist}(\boldsymbol{v}_3, \boldsymbol{v}_2)$.

That is, $\mathbf{dist}(\cdot, \cdot)$ is a pseudo metric on $\mathbb{R}^{\mathscr{N}}$, which explains the name pseudo geometric graph in Definition 3.3.5. $\mathbf{dist}(\cdot, \cdot)$ is indeed a metric on the quotient space of $\mathbb{R}^{\mathscr{N}}$ with any two points symmetric about origin identified.

If the set of points generating geometric graphs or pseudo geometric graphs are random, then the corresponding graphs are called random geometric graphs or random pseudo geometric graphs.

With the above definitions and by the discussions before Definition 3.3.5, the empirical correlation graph $\mathcal{G}_\rho(\boldsymbol{\Psi}^{(\boldsymbol{R})})$ is isomorphic to $\mathbf{PGe}\left(\{\boldsymbol{u}_i\}_{i=1}^p, r_\rho; p, n-1\right)$, the random pseudo geometric graph generated by $\boldsymbol{U}$-scores. Even though $\mathbf{PGe}\left(\{\boldsymbol{u}_i\}_{i=1}^p, r_\rho; p, n-1\right)$ has additional

geometric contents since each vertex in it is a specific point in S^{n-2}, it's not necessary to differentiate it from the empirical correlation graph as long as only the graph properties are of concerned. As an example, we may refer to $N_{V_\delta}^{(\boldsymbol{R})}$ the number of vertices with degree at least δ in $\mathbf{PGe}\left(\{\boldsymbol{u}_i\}_{i=1}^p, r_\rho; p, n-1\right)$ as well. An entirely analogous analysis applies to empirical partial correlation graph and $\mathbf{PGe}\left(\{\boldsymbol{y}_i\}_{i=1}^p, r_\rho; p, n-1\right)$. This equivalent construction indicate the distribution of each of the 3 quantities $\{N_i^{(k)} : i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ with $k = \boldsymbol{R}$ ($k = \boldsymbol{P}$) depends only on the pairwise pseudo distances $\mathbf{dist}(\cdot,\cdot)$ between columns of $\boldsymbol{U}$ ($\boldsymbol{Y}$).

Recall $\mathbf{NMD}\left(\{\boldsymbol{v}_i\}_{i=1}^m, r; m, \mathscr{N}\right)$ denotes the number of vertices of maximum degree $m-1$ in $\mathbf{Ge}\left(\{\boldsymbol{v}_i\}_{i=1}^m, r; m, \mathscr{N}\right)$. When $m \geq 3$, $\mathbf{NMD}\left(\{\boldsymbol{v}_i\}_{i=1}^m, r; m, \mathscr{N}\right)$ is also the number of subgraphs of $\mathbf{Ge}\left(\{\boldsymbol{v}_i\}_{i=1}^m, r; m, \mathscr{N}\right)$ isomorphic to Γ_{m-1}. Define $\mathbf{PNMD}\left(\{\boldsymbol{v}_i\}_{i=1}^m, r; m, \mathscr{N}\right)$ analogously for pseudo geometric graph.

Denote by $\deg(\cdot)$ the degree of a given vertex in the graph. Consider $\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1} \overset{i.i.d.}{\sim} \mathrm{unif}(S^{n-2})$. For $\ell \geq 1$ denote

$$\alpha_{n,\delta}(\ell, r_\rho) := \mathbb{P}\left(\mathbf{PNMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-1\right) = \ell | \deg(\boldsymbol{u}_1') = \delta\right). \tag{3.25}$$

$\alpha_{n,\delta}(\ell, r_\rho)$ depends only on n, δ and the threshold ρ and is abbreviated as $\alpha(\ell, r_\rho)$ when there is no confusion. Moreover, $\alpha(\ell, r_\rho) = 0$ when $\ell \geq \delta + 2$. Define a probability distribution $\boldsymbol{\zeta}_{n,\delta,\rho}$ supported on $[\delta+1]$ with

$$\boldsymbol{\zeta}_{n,\delta,\rho}(\ell) = \frac{\alpha(\ell, r_\rho)/\ell}{\sum_{\ell=1}^{\delta+1}\left(\alpha(\ell, r_\rho)/\ell\right)}. \tag{3.26}$$

Note for the special case $\delta = 1$, by definition $\alpha_{n,1}(2, r_\rho) = 1$ and thus $\boldsymbol{\zeta}_{n,1,\rho}(\ell) = \delta_{\{2\}}$, the Dirac measure at 2.

It will be shown in the next three subsections that the probability distribution $\boldsymbol{\zeta}_{n,\delta,\rho}$ is the distribution of the increment of the compound Poisson that non-asymptotically approximates the $\mathscr{L}(\bar{N}_\delta)$, with $\bar{N}_\delta$ a generic random variable in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$.

3.3.3 Closeness of the distribution of the star subgraph counts to compound Poisson

The first part of this subsection gives the intuition why the distribution of $N_{E_\delta}^{(\boldsymbol{R})}$, the star subgraph counts, approximately is a compound Poisson and derives the associated parameters for the compound Poisson approximation based on random pseudo geometric graph representation developed in Subsection 3.3.2. The second part states a formal proposition that establishes an upper bound of the total variation between $\mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right)$ and the compound Poisson distribution derived in the first part.

Some notations have to be introduced before giving the intuition of why $\mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right)$ approximately is a compound Poisson. Let $SC(r, \boldsymbol{z})$ be the sphere cap with radius r at the center $\boldsymbol{z} \in S^{n-2}$.

Formally,

$$SC(r, \boldsymbol{z}) = \{\boldsymbol{x} \in S^{n-2} : \|\boldsymbol{x} - \boldsymbol{z}\|_2 \leq r\}. \tag{3.27}$$

Define $P_n(r) = \frac{\text{Area}(SC(r,\boldsymbol{z}))}{\text{Area}(S^{n-2})}$, where Area$(\cdot)$ is the area of a subset of S^{n-2}. $P_n(r)$ is the normalized area of the spherical cap with radius r. As is shown in (2.6) in [HR11] [4],

$$P_n(r) = \frac{b_n}{2} \int_{1-\frac{r^2}{2}}^{1} (1-u^2)^{\frac{n-4}{2}} du, \quad \text{when } r \in [0, \sqrt{2}], \tag{3.28}$$

where $b_n = \frac{2\Gamma((n-1)/2)}{\sqrt{\pi}\Gamma((n-2)/2)}$. It follows by simple calculation that

$$P_n(r) = 1 - \frac{\text{Area}(SC(\sqrt{4-r^2}, \boldsymbol{z}))}{\text{Area}(S^{n-2})} = 1 - P_n(\sqrt{4-r^2}) \quad \text{when} \sqrt{2} < r \leq 2,$$

and $P_n(r) = 1$ when $r > 2$. Further properties of $P_n(r)$ are summarized in Lemma 3.14.1.

Denote

$$C_\delta^< := \{\vec{i} = (i_0, i_1, \cdots, i_\delta) \in [p]^{\delta+1} : i_1 < i_2 < \cdots < i_\delta, \text{ and } i_\ell \neq i_0, \ \forall\, \ell \in [\delta]\}. \tag{3.29}$$

For $\vec{i} \in C_\delta^<$, denote by

$$\Phi_{\vec{i}}^{(\boldsymbol{R})} = \prod_{j=1}^{\delta} \Phi_{i_0 i_j}^{(\boldsymbol{R})} = 1\left(\bigcap_{j=1}^{\delta} \{\textbf{dist}(\boldsymbol{u}_{i_0}, \boldsymbol{u}_{i_j}) \leq r_\rho\}\right) \tag{3.30}$$

the indicator that vertex i_0 is connected to each vertex i_j for $j \in [\delta]$ in the empirical correlation graph. Then by definition

$$N_{E_\delta}^{(\boldsymbol{R})} = \sum_{\vec{i} \in C_\delta^<} \Phi_{\vec{i}}^{(\boldsymbol{R})}. \tag{3.31}$$

If $\Phi_{\vec{i}}^{(\boldsymbol{R})}$ for different $\vec{i} \in C_\delta^<$ are independent or weakly dependent, the distribution of $N_{E_\delta}^{(\boldsymbol{R})}$, as a sum of independent or weakly dependent indicator random variables, is expected to approximately be Poisson. This however is not the case since many terms in the summations are highly dependent. Specifically, for any $\vec{i} \in C_\delta^<$, $\Phi_{\vec{i}}^{(\boldsymbol{R})}$ is highly dependent on $\Phi_{\vec{j}}^{(\boldsymbol{R})}$ for any $\vec{j} \in S_{\vec{i}}$ where

$$S_{\vec{i}} := \left\{\vec{j} \in C_\delta^< \backslash \{\vec{i}\} : \bigcup_{\ell=0}^{\delta} \{j_\ell\} = \bigcup_{\ell=0}^{\delta} \{i_\ell\}\right\}. \tag{3.32}$$

[4] [HR11] indeed defines P_0, which is twice of $P_n(r_\rho)$.

$S_{\vec{i}}$ is the set of indexes sharing the same vertices with $\vec{i}$ but with different center and thus $|S_{\vec{i}}| = \delta$. Indeed, provided $\Phi_{\vec{i}}^{(\boldsymbol{R})} = 1$, which is equivalent to $\mathbf{dist}(\boldsymbol{u}_{i_0}, \boldsymbol{u}_{i_j}) \leq r_\rho$ for $\forall j \in [\delta]$,

$$\mathbf{dist}(\boldsymbol{u}_{i_1}, \boldsymbol{u}_{i_j}) \leq \mathbf{dist}(\boldsymbol{u}_{i_1}, \boldsymbol{u}_{i_0}) + \mathbf{dist}(\boldsymbol{u}_{i_0}, \boldsymbol{u}_{i_j}) \leq 2r_\rho, \text{ for } \forall 2 \leq j \leq \delta.$$

That is $\{\boldsymbol{u}_{i_j}\}_{j=2}^{\delta+1}$ are all close to $\boldsymbol{u}_{i_1}$ and hence it's likely there are edges connecting them. In other words it's likely for $\vec{i'} = (i_1, i_0, \cdots, i_\delta)$, $\Phi_{\vec{i'}}^{(\boldsymbol{R})} = 1$.[5] Let

$$U_{\vec{i}} = \sum_{\vec{j} \in S_{\vec{i}}} \Phi_{\vec{j}}^{(\boldsymbol{R})} \tag{3.33}$$

be the sum of highly dependent terms of $\Phi_{\vec{i}}^{(\boldsymbol{R})}$. To sum up, if there is an increment for $N_{E_\delta}^{(\boldsymbol{R})}$, say $\Phi_{\vec{i}}^{(\boldsymbol{R})} = 1$, there is a certain probability that $U_{\vec{i}}$ is great than 0 due to the high dependence, causing each increment of $N_{E_\delta}^{(\boldsymbol{R})}$ has size great than 1 with a certain probability, which is a typical behavior of a compound Poisson distribution.

After understanding heuristically the distribution of $N_{E_\delta}^{(\boldsymbol{R})}$ approximately is a compound Poisson, we now derive heuristically the parameters of the compound Poisson for the special case $\boldsymbol{\Sigma}$ is diagonal. Let $\left[\vec{i}\right] = \{i_0, i_1, \cdots, i_\delta\}$ be the unordered set of indexes of any $\vec{i} \in C_\delta^<$ and define $[C_\delta^<] := \left\{ \left[\vec{i}\right] : \vec{i} \in C_\delta^< \right\}$. It follows $|[C_\delta^<]| = \binom{p}{\delta+1}$. For a given group of $\delta+1$ indexes $\left[\vec{i}\right]$, $\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}$ is the increment associated to this group and its value is between 0 and $\delta+1$. Heuristically the probability of increment size ℓ for $\ell \geq 1$ is proportional to the expectation of the fraction of the number of groups with increment ℓ:[6]

$$\begin{aligned}
\mathbb{E}\frac{1}{|[C_\delta^<]|} \sum_{\left[\vec{i}\right] \in [C_\delta^<]} 1\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} = \ell\right) =& \frac{1}{|[C_\delta^<]|}\frac{1}{\ell}\mathbb{E} \sum_{\left[\vec{i}\right] \in [C_\delta^<]} \left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) 1\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} = \ell\right) \\
=& \frac{1}{|[C_\delta^<]|}\frac{1}{\ell}\mathbb{E} \sum_{\vec{i} \in C_\delta^<} \Phi_{\vec{i}}^{(\boldsymbol{R})} 1\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} = \ell\right) \\
=& \frac{1}{|[C_\delta^<]|}\frac{1}{\ell} \sum_{\vec{i} \in C_\delta^<} \mathbb{P}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} = 1\right) \mathbb{P}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} = \ell | \Phi_{\vec{i}}^{(\boldsymbol{R})} = 1\right).
\end{aligned} \tag{3.34}$$

[5] Here we without loss of generality assume $i_0 < i_j$ for $2 \leq j \leq \delta$.

[6] Here it is implicitly assumed that the random variables with different group of $\delta+1$ indexes are weakly dependent, which will be verified in the proof.

Since $\mathbf{\Sigma}$ is diagonal, $\{\boldsymbol{u}_i\}_{i=1}^p$ are i.i.d. unif(S^{n-2}) and hence (3.34) become

$$\begin{aligned}\mathbb{E}\frac{1}{|[C_\delta^<]|}\sum_{[\vec{i}]\in[C_\delta^<]} 1\left(\Phi_{\vec{i}}^{(\boldsymbol{R})}+U_{\vec{i}}=\ell\right) =& \frac{1}{|[C_\delta^<]|}\frac{1}{\ell}\left|C_\delta^<\right|\left(2P_n(r_\rho)\right)^\delta\alpha(\ell,r_\rho)\\ =& \frac{\delta+1}{\ell}(2P_n(r_\rho))^\delta\alpha(\ell,r_\rho),\end{aligned} \tag{3.35}$$

where $\alpha(\ell,r_\rho)$ is defined in (3.25) and $\mathbb{P}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})}=1\right)=(2P_n(r_\rho))^\delta$ by conditioning on $\boldsymbol{u}_{i_0}$. As a consequence, the probability of increment size ℓ for $\ell\geq 1$ is:

$$\mathbb{E}\frac{1}{|[C_\delta^<]|}\sum_{[\vec{i}]\in[C_\delta^<]} 1\left(\Phi_{\vec{i}}^{(\boldsymbol{R})}+U_{\vec{i}}=\ell\right)\Big/\sum_{\ell=1}^{\delta+1}\left(\mathbb{E}\frac{1}{|[C_\delta^<]|}\sum_{[\vec{i}]\in[C_\delta^<]} 1\left(\Phi_{\vec{i}}^{(\boldsymbol{R})}+U_{\vec{i}}=\ell\right)\right)=\boldsymbol{\zeta}_{n,\delta,\rho}(\ell),$$

where the last step follows from (3.35) and (3.26). The heuristic derivation indicates $\boldsymbol{\zeta}_{n,\delta,\rho}$ is the distribution of increment size of the compound Poisson approximation. Moreover, the mean x of underlying Poisson for the compound Poisson approximation should satisfy the following expectation constraint:

$$\text{expectation of the compound Poisson } = x\mathbb{E}\boldsymbol{\zeta}_{n,\delta,\rho}=\mathbb{E}N_{E_\delta}^{(\boldsymbol{R})},$$

where $\mathbb{E}\boldsymbol{\zeta}_{n,\delta,\rho}$ is the expectation of $\boldsymbol{\zeta}_{n,\delta,\rho}$. One can easily verify $\mathbb{E}\boldsymbol{\zeta}_{n,\delta,\rho}=1/\sum_{\ell=1}^{\delta+1}(\alpha(\ell,r_\rho)/\ell)$ and $\mathbb{E}N_{E_\delta}^{(\boldsymbol{R})}=\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta$ since $\mathbf{\Sigma}$ is diagonal. Hence the mean of underlying Poisson for the compound Poisson is

$$x=\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\sum_{\ell=1}^{\delta+1}\frac{\alpha(\ell,r_\rho)}{\ell}:=\lambda_{p,n,\delta,\rho}. \tag{3.36}$$

In conclusion we heuristically derive when $\mathbf{\Sigma}$ is diagonal, the compound Poisson approximation for $N_{E_\delta}^{(\boldsymbol{R})}$ is CP$(\lambda_{p,n,\delta,\rho},\boldsymbol{\zeta}_{n,\delta,\rho})$.

Despite the above analysis imposes that $\mathbf{\Sigma}$ is diagonal, the general case that $\mathbf{\Sigma}$ is not diagonal but sparse shares the same compound Poisson approximation with a cost of being non-diagonal in the error of the approximation. We are now in a good position to present the main results in this subsection.

Proposition 3.3.6. *Let $p\geq n\geq 4$, $\delta\in[p-1]$ and $\gamma>0$ be given. Suppose $\boldsymbol{X}\sim\mathcal{VE}(\boldsymbol{\mu},\mathbf{\Sigma},\theta)$. Suppose $2p^{1+\frac{1}{\delta}}P_n(r_\rho)\leq\gamma$, and $\mathbf{\Sigma}$ is row-κ sparse. Then*

$$d_{TV}\left(\mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right),CP(\lambda_{p,n,\delta,\rho},\boldsymbol{\zeta}_{n,\delta,\rho})\right)$$

$$\leq C_{n,\delta,\gamma}\left(C'_{\delta,\gamma}\right)^{\mu_{n,\delta+1}(\mathbf{\Sigma})\frac{\kappa-1}{p}}\left(\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\frac{\kappa}{p}\left(1+\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\left(\frac{\kappa}{p}\right)^{2}\right)+p^{-\frac{1}{\delta}}\right),$$

where $C_{n,\delta,\gamma}$ and $C'_{\delta,\gamma}$ are respectively two constants depending only on the parameters in their subscript.

Remark 3.3.7. The condition $2p^{1+\frac{1}{\delta}}P_n(r_\rho)\leq\gamma$ specifies an implicit lower bound on the threshold ρ. To obtain an explicit lower bound, observe $2c_np^{1+\frac{1}{\delta}}\left(\sqrt{2(1-\rho)}\right)^{n-2}\leq\gamma$ is a sufficient condition of $2p^{1+\frac{1}{\delta}}P_n(r_\rho)\leq\gamma$, by Lemma 3.14.1 (a). Solving for ρ, then an explicit lower bound of ρ sufficient for $2p^{1+\frac{1}{\delta}}P_n(r_\rho)\leq\gamma$ is

$$\rho\geq 1-\frac{1}{2}\left(\frac{\gamma}{2c_np^{1+\frac{1}{\delta}}}\right)^{\frac{2}{n-2}}.$$

This is a non-asymptotic version of (3.12).

Even though Proposition 3.3.6 holds for any symmetric positive definite matrix $\mathbf{\Sigma}$ that is row-κ sparse, for the results to be effective, the upper bounds in the above results should be small. As a result, $\mathbf{\Sigma}$ has to have relatively small $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)$ and be row-κ sparse with relative small sparsity level κ/p, such that $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\kappa/p$ is small. A sufficient condition by Lemma 3.7.1 (b) is small condition number of $\mathbf{\Sigma}$ to guarantee small $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)$. In the special case when $\mathbf{\Sigma}$ is diagonal, $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\kappa/p=1/p$.

Moreover, suppose $\mathbf{\Sigma}$ has small condition number and sparsity level such that $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\kappa/p$ is small, say $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\kappa/p<1$. Then in the upper bound the term $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\left(\kappa/p\right)^2$ inside the parenthesis can be dropped, resulting in an additional constant factor, since $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\left(\kappa/p\right)^2<1$. In other words the effective upper bound, neglecting the coefficients depending on n, δ and γ, is $\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\kappa/p+p^{-\frac{1}{\delta}}$.

The expressions for $C'_{\delta,\gamma}$ and $C_{n,\delta,\rho}$ are respectively available in (3.89) and (3.90). They are not optimal constants since they are not of major concern in this chapter. Here possible expressions for these coefficients are provided for completeness. □

Proposition 3.3.6 states for given n, p, δ and γ, if the threshold ρ is properly chosen, and $\mathbf{\Sigma}$ is row-κ sparse and has small $\mu_{n,2\delta+2}(\mathbf{\Sigma})$, then the distribution of $N_{E_\delta}^{(\boldsymbol{R})}$ approximately is the compound Poisson $CP(\lambda_{p,n,\delta,\rho},\boldsymbol{\zeta}_{n,\delta,\rho})$. In the next subsection, we will built connections among $\{N_i^{(k)}:k\in\{\boldsymbol{R},\boldsymbol{P}\},i\in\{E_\delta,\breve{V}_\delta,V_\delta\}\}$ such that Proposition 3.3.6 can be extended to the other 5 quantities.

3.3.4 A portmanteau proposition on pairwise total variations

In this subsection upper bounds for pairwise total variation distances among the 6 quantity $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ are obtained. Such an result establishes the distribution of these 6 quantities are mutually close, and together with Proposition 3.3.6 it will imply their distributions all are close to the compound Poisson $CP(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$. Like in Subsection 3.3.3, some intuitive derivations are first presented, followed by a formal statement of the results.

Lemma 3.3.8. *Consider* $\delta \in [p-2]$.

$$N_{E_\delta}^{(\boldsymbol{R})} - (\delta+1)N_{E_{\delta+1}}^{(\boldsymbol{R})} \le N_{\breve{V}_\delta}^{(\boldsymbol{R})} \le N_{V_\delta}^{(\boldsymbol{R})} \le N_{E_\delta}^{(\boldsymbol{R})},$$

$$N_{E_\delta}^{(\boldsymbol{P})} - (\delta+1)N_{E_{\delta+1}}^{(\boldsymbol{P})} \le N_{\breve{V}_\delta}^{(\boldsymbol{P})} \le N_{V_\delta}^{(\boldsymbol{P})} \le N_{E_\delta}^{(\boldsymbol{P})}.$$

It follows directly from Lemma 3.3.8 that for $\tilde{N}_\delta \in \left\{N_{\breve{V}_\delta}^{(\boldsymbol{R})}, N_{V_\delta}^{(\boldsymbol{R})}\right\}$,

$$\mathbb{E}\left|\tilde{N}_\delta - N_{E_\delta}^{(\boldsymbol{R})}\right| \le (\delta+1)\mathbb{E}N_{E_{\delta+1}}^{(\boldsymbol{R})}. \tag{3.37}$$

As a result, if $\mathbb{E}N_{E_{\delta+1}}^{(\boldsymbol{R})}$ is small, then $N_{\breve{V}_\delta}^{(\boldsymbol{R})}$ and $N_{V_\delta}^{(\boldsymbol{R})}$ are close to $N_{E_\delta}^{(\boldsymbol{R})}$ in L^1 norm.

To heuristically see why the quantities in empirical partial correlation graph is close to those in the empirical correlation graph, consider large p and pretend $\{\boldsymbol{u}_i\}_{i=1}^p$ are independent. Then according to law of large number,

$$\boldsymbol{B} = \frac{n-1}{p}\sum_{i=1}^p \boldsymbol{u}_i\boldsymbol{u}_i^T \approx (n-1)\mathbb{E}\boldsymbol{u}_i\boldsymbol{u}_i^T = \boldsymbol{I}_{n-1}, \tag{3.38}$$

which further implies

$$\bar{\boldsymbol{Y}} \approx \boldsymbol{U}, \quad \boldsymbol{Y} \approx \boldsymbol{U}.$$

That is, the $\boldsymbol{Y}$-score and $\boldsymbol{U}$-score are almost the same. Hence $N_{E_\delta}^{(\boldsymbol{R})}$ and $N_{E_\delta}^{(\boldsymbol{P})}$, as the same function of $\boldsymbol{U}$ and $\boldsymbol{Y}$ respectively, are close to each other. So does the other two pairs. These heuristic arguments will be made rigorous in the proof of the next result.

The next result states all 6 quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ are close to each other in L^1 norm and their distribution are then close to each other in total variation.

Proposition 3.3.9. *Let* $p \ge n \ge 4$ *and* $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. *Let* $\delta \in [p-1]$. *Suppose* $2p^{1+\frac{1}{\delta}}P_n(r_\rho) \le \gamma$.

(a) Suppose $\mathbf{\Sigma}$ is row-κ sparse. Then for $\tilde{N}_\delta \in \left\{N_{\check{V}_\delta}^{(\boldsymbol{R})}, N_{V_\delta}^{(\boldsymbol{R})}\right\}$,

$$d_{TV}\left(\mathscr{L}\left(\tilde{N}_\delta\right), \mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right)\right) \leq \mathbb{E}\left|\tilde{N}_\delta - N_{E_\delta}^{(\boldsymbol{R})}\right| \leq \frac{(\delta+1)^2}{\delta!}\gamma^{\delta+1}\left(1+\mu_{n,\delta+2}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right)p^{-\frac{1}{\delta}}.$$

(b) Suppose $\mathbf{\Sigma}$, after some row-column permutation, is (τ,κ) sparse with $\tau \leq \frac{p}{2}$ and

$$\left(\sqrt{\frac{n-1}{p}} + \sqrt{\frac{\delta \ln p}{p}}\right) \leq c$$

hold for some positive and small universal constant c. Then

$$\begin{aligned} d_{TV}\left(\mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{P})}\right), \mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right)\right) \leq & \mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{P})} - N_{E_\delta}^{(\boldsymbol{R})}\right| \\ \leq & C_{E_\delta}^{(\boldsymbol{P})}\left(1+\frac{\kappa-1}{p}\mu_{n,\delta+1}(\mathbf{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}}+\frac{\tau}{p}\right), \end{aligned} \tag{3.39}$$

where $C_{E_\delta}^{(\boldsymbol{P})}$ is a constant depending only on n, δ and γ.

(c) Suppose the same conditions as in part (b) hold. Then

$$\begin{aligned} d_{TV}\left(\mathscr{L}\left(N_{\check{V}_\delta}^{(\boldsymbol{P})}\right), \mathscr{L}\left(N_{\check{V}_\delta}^{(\boldsymbol{R})}\right)\right) \leq & \mathbb{E}\left|N_{\check{V}_\delta}^{(\boldsymbol{P})} - N_{\check{V}_\delta}^{(\boldsymbol{R})}\right| \\ \leq & C_{\check{V}_\delta}^{(\boldsymbol{P})}\left(1+\frac{\kappa-1}{p}\mu_{n,\delta+2}(\mathbf{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}}+\frac{\tau}{p}+p^{-\frac{1}{\delta}}\right), \end{aligned} \tag{3.40}$$

where $C_{\check{V}_\delta}^{(\boldsymbol{P})}$ is a constant depending only on n, δ and γ.

Remark 3.3.10. In Proposition 3.3.9 (b), the condition

$$\left(\sqrt{\frac{n-1}{p}} + \sqrt{\frac{\delta \ln p}{p}}\right) \leq c \tag{3.41}$$

hold for some positive and small enough universal constant c is nothing but a quantitative way of saying when p is sufficiently large. Observe the left side of (3.41) is a decreasing function of p, and its limit is 0 when p approaches infinity. Then the smallest positive integer p_0 satisfying the inequality exists and depends only on n and δ, since c is an universal constant. Then (3.41) is equivalent to requiring $p \geq p_0$. Similar interpretation applies for the corresponding condition in Proposition 3.3.9 (c).

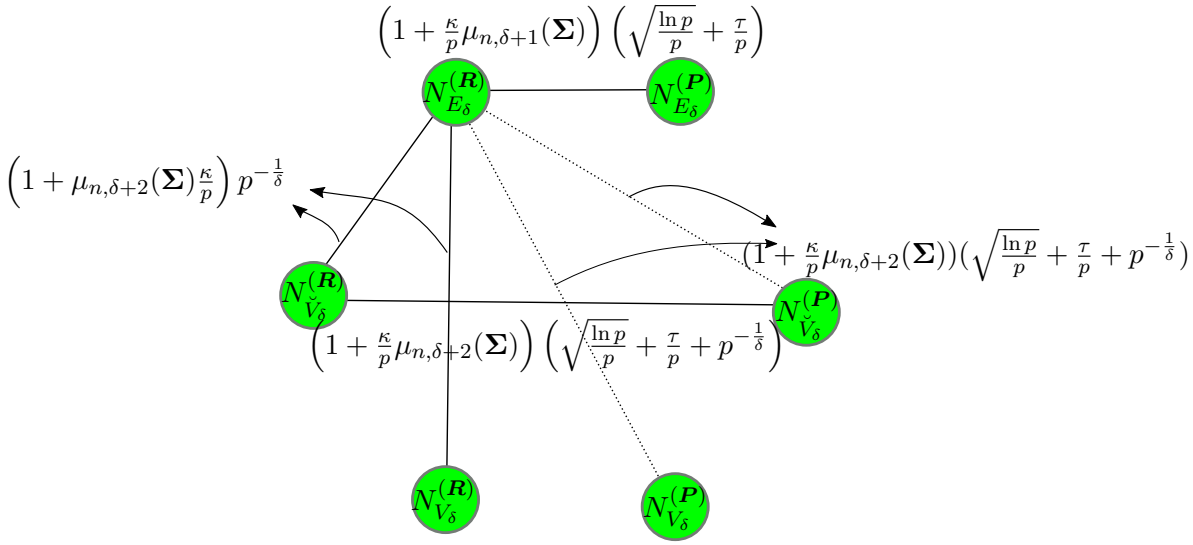


Figure 3.3: This graph has the 6 quantities associated to empirical correlation or partial correlation graph as vertices. The 4 solid edges correspond to existence of an direct upper bound of the total variation between two vertices, with the weights respectively correspond to the 4 upper bounds (neglecting constant coefficients) in Proposition 3.3.9. Dash edges correspond to an indirect upper bound of the total variation between vertices, with weights computed from solid path connecting the two vertices.

Row-κ sparsity on $\mathbf{\Sigma}$ suffices to show the quantities of empirical correlation graph are close in L^1 norm as in Proposition 3.3.9 (a). Stronger sparse condition (τ,κ) sparsity on $\mathbf{\Sigma}$ is imposed to have quantities between empirical correlation graph and empirical partial correlation graph are close in L^1 norm as in Proposition 3.3.9 (b), (c). (τ,κ) sparsity is indeed only used to guarantee (3.38) and have a quantitative control how $\boldsymbol{B}$ deviates from $\boldsymbol{I}_{n-1}$.

Even though Proposition 3.3.9 holds for any symmetric positive definite matrix $\mathbf{\Sigma}$ that, after simultaneous row-column permutation, is (τ,κ) sparse, for the results to be effective, the upper bounds in the proposition should be small. All 3 upper bounds in the proposition, up to a constant depending on n, δ and γ, are bounded by $\left(1+\frac{\kappa-1}{p}\mu_{n,\delta+2}(\mathbf{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}}+\frac{\tau}{p}+p^{-\frac{1}{\delta}}\right)$. As a result, for the proposition to be useful the theoretical covariance matrix $\mathbf{\Sigma}$ should has small $\mu_{n,\delta+2}(\mathbf{\Sigma})$ and be (τ,κ) sparse with small sparsity level $\frac{\tau}{p}$, $\frac{\kappa}{p}$, and p should be relatively large such that $\left(1+\frac{\kappa-1}{p}\mu_{n,\delta+2}(\mathbf{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}}+\frac{\tau}{p}+p^{-\frac{1}{\delta}}\right)$ is small. A sufficient condition by Lemma 3.7.1 (b) is small condition number of $\mathbf{\Sigma}$ to guarantee small $\mu_{n,\delta+2}\left(\mathbf{\Sigma}\right)$.

The exact expressions for $C^{(\boldsymbol{P})}_{E_\delta}$ and $C^{(\boldsymbol{P})}_{\check{V}_\delta}$ are available respectively at (3.109) and (3.114). They are not optimal constants since they are not of major concern in this chapter. Here possible expressions for these coefficients are provided for completeness.

□

Proposition 3.3.9 establishes the total variation bounds between $N^{(k)}_{E_\delta}, N^{(k)}_{\check{V}_\delta}, N^{(k)}_{V_\delta}$ with $k=\boldsymbol{R}$ and $k=\boldsymbol{P}$ as illustrated by Figure 3.3. Dash edges correspond to an indirect upper bound of the

total variation between vertices, with weights computed from solid path connecting the two vertices. For instance the weight of dash edge between $N_{E_\delta}^{(\boldsymbol{R})}$ and $N_{V_\delta}^{(\boldsymbol{P})}$ is computed from

$$\begin{aligned}
&d_{\text{TV}}\left(\mathscr{L}\left(N_{V_\delta}^{(\boldsymbol{P})}\right), \mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right)\right)\\
\leq&\mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{R})} - N_{V_\delta}^{(\boldsymbol{P})}\right|\\
\leq&\mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{R})} - N_{\breve{V}_\delta}^{(\boldsymbol{P})}\right| + \mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{R})} - N_{E_\delta}^{(\boldsymbol{P})}\right|\\
\leq&\mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{R})} - N_{\breve{V}_\delta}^{(\boldsymbol{R})}\right| + \mathbb{E}\left|N_{\breve{V}_\delta}^{(\boldsymbol{R})} - N_{\breve{V}_\delta}^{(\boldsymbol{P})}\right| + \mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{R})} - N_{E_\delta}^{(\boldsymbol{P})}\right|\\
\leq&\left(C_{E_\delta}^{(\boldsymbol{P})} + C_{\breve{V}_\delta}^{(\boldsymbol{P})} + \frac{(\delta+1)^2}{\delta!}\gamma^{\delta+1}\right)\left(1 + \frac{\kappa-1}{p}\mu_{n,\delta+2}(\boldsymbol{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}} + \frac{\tau}{p} + p^{-\frac{1}{\delta}}\right),
\end{aligned}$$

where the first inequality follows from Lemma 3.14.4, the second inequality follows from Lemma 3.3.8, and the last inequality follows from Proposition 3.3.9 (a), (b) and (c).

From Proposition 3.3.9 and Figure 3.3, it's easy to see by triangle inequality the 6 quantities are all close to each other in total variation provided $\left(1 + \frac{\kappa-1}{p}\mu_{n,\delta+2}(\boldsymbol{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}} + \frac{\tau}{p} + p^{-\frac{1}{\delta}}\right)$ is relatively small. As a result, the closeness of one quantity among the 6 to some distribution in total variation implies the closeness of all 6 quantities to that distribution. In Subsection 3.3.3 the result that $\mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right)$ is close to the compound Poisson $CP(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ has been established, which immediately implies all 6 quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ are close in total variation to the same compound Poisson. A formal such result combining Proposition 3.3.6 and Proposition 3.3.9 is presented in next subsection.

3.3.5 Unified convergence: an umbrella theorem

The following theorem is a non-asymptotic version of Theorem 3.2.4. It states if the threshold ρ is properly chosen, and $\boldsymbol{\Sigma}$ satisfies (τ, κ) sparsity condition, then any random quantity in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ can be approximated by a compound Poisson.

Theorem 3.3.11 (Compound Poisson Approximation in High Dimension)**.** *Let $n \geq 4$, $\delta \in [p-1]$, and $\gamma > 0$ be given. Consider $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $2p^{1+\frac{1}{\delta}}P_n(r_\rho) \leq \gamma$. Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \leq \frac{p}{2}$ and $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})\frac{\kappa}{p} < 1$. Suppose $\sqrt{\frac{n-1}{p}} + \sqrt{\frac{\delta \ln p}{p}} \leq c$ hold for some positive and small universal constant c. Let $\bar{N}_\delta$ be a generic random variable for either one in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$. Then*

$$d_{TV}\left(\mathscr{L}\left(\bar{N}_\delta\right), CP(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})\right) \leq C_{n,\delta,\gamma}\left(\mu_{n,2\delta+2}\left(\boldsymbol{\Sigma}\right)\frac{\kappa}{p} + p^{-\frac{1}{\delta}} + E(p,\delta)\right), \tag{3.42}$$

where

$$E(p,\delta)=\begin{cases}0 & \textit{if } \bar{N}_\delta = N_{E_\delta}^{(\boldsymbol{R})},\ N_{\breve{V}_\delta}^{(\boldsymbol{R})} \textit{ or } N_{V_\delta}^{(\boldsymbol{R})},\\ \sqrt{\frac{\ln p}{p}}+\frac{\tau}{p} & \textit{if } \bar{N}_\delta = N_{E_\delta}^{(\boldsymbol{P})}, N_{\breve{V}_\delta}^{(\boldsymbol{P})} \textit{ or } N_{V_\delta}^{(\boldsymbol{P})}.\end{cases}$$

Remark 3.3.12 (Relaxed sparsity assumption in empirical correlation graph)**.** If only random quantities in empirical correlation graph is of concern, then the (τ,κ) sparsity assumption can be relaxed to the row-κ sparsity. Specifically, the last three sentences in Theorem 3.3.11 can be replaced by the following.

Suppose $\boldsymbol{\Sigma}$ is row-κ sparse with $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})\frac{\kappa}{p}<1$. Let $\tilde{N}_\delta$ be a generic random variable for either one in $\{N_i^{(k)} : k=\boldsymbol{R}, i\in\{E_\delta,\breve{V}_\delta,V_\delta\}\}$. Then

$$d_{\mathrm{TV}}\left(\mathscr{L}\left(\tilde{N}_\delta\right), CP(\lambda_{p,n,\delta,\rho},\boldsymbol{\zeta}_{n,\delta,\rho})\right)\le C_{n,\delta,\gamma}\left(\mu_{n,2\delta+2}\left(\boldsymbol{\Sigma}\right)\frac{\kappa}{p}+p^{-\frac{1}{\delta}}\right), \tag{3.43}$$

where the notation $C_{n,\delta,\gamma}$ is a constant depending on n, δ and γ.

Remark 3.3.13. The condition $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})\frac{\kappa}{p}<1$ is used to simplify the upper bound in (3.43) and (3.42), in the sense that without this condition, inequalities similar to (3.43) and (3.42) but with more complicated upper bounds still hold. This condition is not really an additional condition: observe for upper bound in (3.43) and (3.42) (neglecting the coefficients depending only on n,δ, and γ) to be small, the term $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})\frac{\kappa}{p}$ should be small. See the third paragraph in Remark 3.3.7 for a detailed discussion.

Row-κ sparsity on $\boldsymbol{\Sigma}$ suffices to guarantee (3.43) holds, i.e. the quantities of interest in the empirical correlation graph can be approximated by a compound Poisson. For Remark 3.3.12 to be useful, the upper bound (3.43) should be small such that the distribution of $\tilde{N}_\delta$ is close to $CP(\lambda_{p,n,\delta,\rho},\boldsymbol{\zeta}_{n,\delta,\rho})$ can be drawn. Neglecting the coefficient depending on n,δ, and γ since n, δ and γ are given, it suffices to have the term $\left(\mu_{n,2\delta+2}(\boldsymbol{\Sigma})\frac{\kappa}{p}+p^{-\frac{1}{\delta}}\right)$ be small. That is, for the above theorem to be effective, $\boldsymbol{\Sigma}$ should has small $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})$, small sparsity level $\frac{\kappa}{p}$, and p should be relatively large. One sufficient condition for small $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})$ is small condition number of $\boldsymbol{\Sigma}$ by Lemma 3.7.1 (b).

The stronger condition (τ,κ) sparsity on $\boldsymbol{\Sigma}$ is imposed to guarantee (3.42) holds, i.e. the quantities of interest in empirical partial correlation graph can be approximated by the same compound Poisson as that in empirical correlation graph. For Theorem 3.3.11 to be useful, $\boldsymbol{\Sigma}$ should has small $\mu_{n,2\delta+2}(\boldsymbol{\Sigma})$, small sparsity level $\frac{\tau}{p}$, $\frac{\kappa}{p}$, and p should be relatively large.

Finally observe (3.43) and (3.42) do not involve the parameters $\boldsymbol{\mu}$ and θ. That is, the above theorem holds for any mean $\boldsymbol{\mu}$ and any shaping function θ of the distribution of the data matrix $\boldsymbol{X}$. The reason for this observation has been discussed in Subsection 3.3.1. □

Remark 3.3.14 (Comparisons between theorems)**.** Theorem 3.3.11 is a non-asymptotic version compound Poisson approximation and Theorem 3.2.4 is the limiting version. Note by taking the limit $p \to \infty$, we obtain simpler formulae for parameters of the corresponding compound Poisson. Specifically the distribution of increment $\boldsymbol{\zeta}_{n,\delta,\rho}$ of the approximating compound Poisson in Theorem 3.3.11 depends on conditional probabilities in random pseudo geometric graph as in (3.25). On the other hand, the distribution of increment $\boldsymbol{\zeta}_{n,\delta}$ of the limiting compound Poisson in Theorem 3.2.4 depends on probabilities in random geometric graph as in (3.7), which is relatively simpler. For instance, a closed form formula of $\boldsymbol{\zeta}_{n,2}$ is obtained later in Example 3.5.2 while obtaining a closed formula for $\boldsymbol{\zeta}_{n,2,\rho}$ does not seem straightforward.

Despite the fact that the limiting compound Poisson in Theorem 3.2.4 is relatively simpler, the disadvantage of it is that it requires that $\rho \to 1$ in the specific rate $p^{-\frac{2}{n-2}\left(1+\frac{1}{\delta}\right)}$ as discussed in Remark 3.2.6. This particular rate, however, is very slow when n is large. Indeed if one choose ρ as in (3.12) and require $\rho \geq 1-\epsilon$ for some $\epsilon \in (0, 1/2)$, one obtains

$$p \geq \left(\frac{e_{n,\delta}}{2c_n}\right)^{\frac{1}{1+1/\delta}} \left(\frac{1}{2\epsilon}\right)^{\frac{n-2}{2(1+1/\delta)}}.$$

It is clear from the preceding display that when n is large, p is huge. On the contrary, Theorem 3.3.11 does not impose the requirement that ρ approach 1 and approximates the $\bar{N}_\delta$ even for small p. This is illustrated in Figure 3.8 in Section 3.15.

Another advantage of Theorem 3.3.11 is that explicit upper bounds for the approximation errors are established, while only limiting results but no convergence rates are established in Theorem 3.2.4. Though from the discussion in the previous paragraph, one should expect the convergence rate for Theorem 3.2.4 to be slow for large n. □

Theorem 3.3.11 and Remark 3.3.12 directly follow from Proposition 3.3.6 and Proposition 3.3.9 and hence the proof is omitted. In the rest of this subsection we present results on the limit of $\text{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ when p goes to infinity and then complete the proof of Theorem 3.2.4.

To study the limit distribution for $\text{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$, the following two results in random geometric graphs are useful.

The next proposition states that the distribution of the number of vertices of maximum degree, conditioned on the existence of one such vertex, is invariant in geometric graph and pseudo geometric graph generated by vertices uniformly distributed on the unit sphere.

Lemma 3.3.15. *Consider $r < 2/\sqrt{5}$ and $\delta \geq 1$. Suppose $\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1} \overset{i.i.d.}{\sim}$ unif(S^{n-2}). Then for any $\ell \in [\delta+1]$,*

$$\mathbb{P}\left(\boldsymbol{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r; \delta+1, n-1\right) = \ell | \boldsymbol{deg}(\boldsymbol{u}_{\delta+1}') = \delta\right)$$

$$=\mathbb{P}\left(\boldsymbol{PNMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r; \delta+1, n-1\right)=\ell | \boldsymbol{deg}(\boldsymbol{u}_{\delta+1}')=\delta\right).$$

where $\boldsymbol{deg}(\boldsymbol{u}_{\delta+1}')$ *on the left (right) side is the degree of vertex* $\boldsymbol{u}_{\delta+1}'$ *in the corresponding random (pseudo) geometric graph.*

By Lemma 3.3.15,

$$\alpha_{n,\delta}(\ell, r_\rho)=\mathbb{P}\left(\mathbf{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-1\right)=\ell | \text{deg}(\boldsymbol{u}_1')=\delta\right) \tag{3.44}$$

when $r_\rho < 2/\sqrt{5}$ or equivalently $\rho > 3/5$.

Lemma 3.3.16. *Let* $\delta \geq 1$ *and* $n \geq 3$. *Suppose* $\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1} \overset{i.i.d.}{\sim}$ *unif*(S^{n-2}) *and* $\{\tilde{\boldsymbol{u}}\}_{i=1}^{\delta} \overset{i.i.d.}{\sim}$ *unif*(B^{n-2}). *Then for any* $\ell \in [\delta+1]$,

$$\begin{aligned}&\lim_{r\to 0^+}\mathbb{P}\left(\boldsymbol{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r; \delta+1, n-1\right)=\ell | \boldsymbol{deg}(\boldsymbol{u}_{\delta+1}')=\delta\right)\\ =&\mathbb{P}\left(\boldsymbol{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2\right)=\ell-1\right).\end{aligned} \tag{3.45}$$

Lemma 3.3.16 states that the conditional distribution of the number of vertices of maximum degree in $\mathbf{Ge}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r; \delta+1, n-1\right)$, conditioned on the existence of one such vertex, convergences. Its limit is the distribution of number of vertices of maximum degree in $\mathbf{Ge}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2\right)$, the random geometric graph generated by uniform distribution in the unit ball.

The condition $2p^{1+\frac{1}{\delta}}P_n(r_\rho) \leq \gamma$ in Theorem 3.3.11 entails $r_\rho \to 0^+$, which is equivalent to $\rho \to 1^-$, when $p \to \infty$. The following lemma states if the rate of $\rho \to 1^-$ is coupled with the rate $p \to \infty$, $\text{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho}) \to \text{CP}(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta})$ in distribution, where $\boldsymbol{\zeta}_{n,\delta}$ is defined (3.8) and $\lambda_{n,\delta}(e_{n,\delta})$ is defined in Theorem 3.2.4.

Lemma 3.3.17. *Suppose as* $p \to \infty$, $\rho \to 1^-$ *such that* $c_n 2^{\frac{n}{2}} p^{1+\frac{1}{\delta}}(1-\rho)^{\frac{n-2}{2}} \to e_{n,\delta}$, *where* $c_n = \frac{\Gamma((n-1)/2)}{(n-2)\sqrt{\pi}\Gamma((n-2)/2)}$ *and* $e_{n,\delta}$ *is some positive constant that possibly depends on* n *and* δ. *Then*

$$CP(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho}) \to CP(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta}) \text{ in distribution}. \tag{3.46}$$

A formal theorem summarizing the results when $p \to \infty$ has been presented in Theorem 3.2.4.

Proof of Theorem 3.2.4: It directly follows from Theorem 3.3.11 and Lemma 3.3.17. □

Note Remark 3.2.5 directly follows from Remark 3.3.12 and Lemma 3.3.17.

3.4 Convergence of moments

Despite that we have shown in Theorem 3.2.4 that $\bar{N}_\delta \to CP(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta})$ in distribution where $\bar{N}_\delta$ is a generic random variable in the set $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$, the

convergence of moments remains unknown. Convergence in distribution does not necessarily induce convergence of moments: there exists $\{0,1\}$-valued random variables Z_n that converges to 0 in distribution but not in first moment (c.f. Example 4.5 in Chapter 3 of [Cin11]). The analogous analysis applies to the non-asymptotic approximation result Theorem 3.3.11. However, convergence or non-asymptotic approximation of moments is important. For instance, in [HR11] approximation formula of the first moment is used to derive a phase transition threshold by $d\mathbb{E}[\bar{N}_\delta]/d\rho = -1$. In this subsection, we present the non-asymptotic approximation of the first moment and second moment of $\bar{N}_\delta$, which will automatically imply convergence results when $p \to \infty$.

Let $Z \sim \mathrm{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$. Then we can represent $Z = \sum_{i=1}^{N} Z_i$, where N is distributed as a Poisson with mean $\lambda_{p,n,\delta,\rho}$, $Z_i \overset{i.i.d.}{\sim} \boldsymbol{\zeta}_{n,\delta,\rho}$ and N is independent of each Z_i. The first two moments of Z are:

$$\begin{aligned}\mathbb{E}Z =& \mathbb{E}N\mathbb{E}Z_1 = \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta, &(3.47)\\ \mathbb{E}Z^2 =& \mathbb{E}N\mathbb{E}Z_1^2 + (\mathbb{E}N\mathbb{E}Z_1)^2 \\ =& \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho) + \left(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\right)^2. &(3.48)\end{aligned}$$

The next lemma is on non-asymptotic approximation of the first moment of $N_{E_\delta}^{(\boldsymbol{R})}$ by the first moment of the compound Poisson as in (3.47).

Lemma 3.4.1. *Let $p \geq n \geq 4$, $\delta \in [p-1]$ and $\gamma > 0$ be given. Suppose $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $2p^{1+\frac{1}{\delta}} P_n(r_\rho) \leq \gamma$, and $\boldsymbol{\Sigma}$ is row-κ sparse. Then*

$$\left|\mathbb{E}N_{E_\delta}^{(\boldsymbol{R})} - \mathbb{E}Z\right| \leq \frac{(\delta+1)}{2((\delta-1)!)}\gamma^\delta \mu_{n,\delta+1}(\boldsymbol{\Sigma})\frac{\kappa-1}{p},$$

where $Z \sim \mathrm{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ and $\mathbb{E}Z$ is calculated in (3.47).

By combining the preceding lemma and Proposition 3.3.9 one immediately obtains non-asymptotic approximations for the first moment of all 6 quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$. If further impose $c_n 2^{\frac{n}{2}} p^{1+\frac{1}{\delta}}(1-\rho)^{\frac{n-2}{2}} \to e_{n,\delta}$ as $p \to \infty$ as in Theorem 3.2.4, then one also obtains a limit version on first moment. All these straightforward extensions are left to the interested readers.

The approximation to second moment of $N_{E_\delta}^{(\boldsymbol{R})}$ involves approximations to terms $\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}$ for $\vec{i}, \vec{j} \in C_\delta^<$, which are already available in the proof of Proposition 3.3.6 when Stein's method is applied. By those results and careful analysis, the approximation of second moment of $N_{E_\delta}^{(\boldsymbol{R})}$ is as below.

Proposition 3.4.2. *Let $p \geq n \geq 4$, $\delta \in [p-1]$ and $\gamma > 0$ be given. Suppose $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $2p^{1+\frac{1}{\delta}} P_n(r_\rho) \leq \gamma$, and $\boldsymbol{\Sigma}$ is row-κ sparse. Then*

$$\left| \mathbb{E}\left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2 - \mathbb{E}Z^2 \right| \leq C_{n,\delta,\gamma} \left(\mu_{n,2\delta+2}(\boldsymbol{\Sigma}) \frac{\kappa}{p} + p^{-1/\delta} \right),$$

where $Z \sim \mathrm{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ and $\mathbb{E}Z^2$ is calculated in (3.48)*.*

To extend the preceding proposition to second moments of other quantities in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$, one need to generalize Proposition 3.3.9 to L^1 distance between the square of the random quantities.

Proposition 3.4.3. *Let $p \geq n \geq 4$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Let $\delta \in [p-1]$. Suppose $2p^{1+\frac{1}{\delta}} P_n(r_\rho) \leq \gamma$.*

(a) Suppose $\boldsymbol{\Sigma}$ is row-κ sparse. Then for $\tilde{N}_\delta \in \left\{ N_{\breve{V}_\delta}^{(\boldsymbol{R})}, N_{V_\delta}^{(\boldsymbol{R})} \right\}$,

$$\mathbb{E}\left| \left(\tilde{N}_\delta\right)^2 - \left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2 \right| \leq C_{n,\delta,\gamma} \left(1 + \mu_{n,2\delta+3}(\boldsymbol{\Sigma}) \frac{\kappa - 1}{p} \right) p^{-1/\delta}.$$

(b) Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \leq \frac{p}{2}$ and

$$\left(\sqrt{\frac{n-1}{p}} + \sqrt{\frac{\delta \ln p}{p}} \right) \leq c$$

hold for some positive and small universal constant c. Then

$$\mathbb{E}\left| \left(N_{E_\delta}^{(\boldsymbol{P})}\right)^2 - \left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2 \right| \leq C_{n,\delta,\gamma} \left(1 + \mu_{n,2\delta+2}(\boldsymbol{\Sigma}) \frac{\kappa - 1}{p} \right) \left(\frac{\sqrt{\ln p}}{\sqrt{p}} + \frac{\tau}{p} \right).$$

(c) Suppose the same conditions as in part (b) hold. Then for $\tilde{N}_\delta \in \left\{ N_{\breve{V}_\delta}^{(\boldsymbol{P})}, N_{V_\delta}^{(\boldsymbol{P})} \right\}$

$$\mathbb{E}\left| \left(\tilde{N}_\delta\right)^2 - \left(N_{E_\delta}^{(\boldsymbol{P})}\right)^2 \right| \leq C_{n,\delta,\gamma} \left(1 + \mu_{n,2\delta+3}(\boldsymbol{\Sigma}) \frac{\kappa - 1}{p} \right) p^{-1/\delta}.$$

By applying triangle inequalities to the preceding proposition, one obtain for $\tilde{N}_\delta \in \{N_{V_\delta}^{(\boldsymbol{P})}, N_{\breve{V}_\delta}^{(\boldsymbol{P})}\}$

$$\mathbb{E}\left| \left(\tilde{N}_\delta\right)^2 - \left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2 \right| \leq C_{n,\delta,\gamma} \left(1 + \mu_{n,2\delta+3}(\boldsymbol{\Sigma}) \frac{\kappa - 1}{p} \right) \left(\frac{\sqrt{\ln p}}{\sqrt{p}} + \frac{\tau}{p} + p^{-1/\delta} \right).$$

Thus we have established the L^1 distance between the square of each term in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ and $\left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2$.

By combining Proposition 3.4.2 and Proposition 3.4.3 one immediately obtains non-asymptotic approximations for all 6 quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$. If further impose $c_n 2^{\frac{n}{2}} p^{1+\frac{1}{\delta}} (1-\rho)^{\frac{n-2}{2}} \to e_{n,\delta}$ as $p \to \infty$ as in Theorem 3.2.4, then one also obtains a limit version on second moment. All these straightforward extensions are left to the interested readers.

For general higher moments the answer remain unknown. One possible direction is to prove that the sequence $\left(\bar{N}_\delta\right)^m$ indexed by p is uniformly integrable, where $\bar{N}_\delta$ is a generic random variable in $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$. Then convergence of s-th moments with $s < m$ follows by Proposition 5.10 in [Cin11] and convergence in distribution established in Theorem 3.2.4.

3.5 Explicit characterizations

The limiting compound Poisson in Theorem 3.2.4 is defined in terms of α_ℓ, while the non-asymptotic compound Poisson in Theorem 3.3.11 is defined in terms of $\alpha(\ell, r_\rho)$. Moreover the second moment approximation established in Proposition 3.4.2 also involve the term $\alpha(\ell, r_\rho)$ due to (3.48). However α_ℓ and $\alpha(\ell, r_\rho)$ are quantities in random geometric graphs, which might not be easy to compute. In this section, we obtain closed-form expressions for $\alpha(\ell, r_\rho)$ and $\alpha(\ell, r_\rho)$ for small δ and provide approximation for them for large n and δ, which implies that the compound Poisson is approximately a Poisson for large n and δ. We shall begin in Subsection 3.5.1 with the study of simpler quantity α_ℓ, and then study $\alpha(\ell, r_\rho)$ in Subsection 3.5.2.

3.5.1 Explicit characterizations for α_ℓ

The limiting compound Poisson CP$(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta})$ has parameters given in (3.8) and in Theorem 3.2.4. However, formulae for underlying Poisson rate $\lambda_{n,\delta}(e_{n,\delta})$ and the distribution of increment $\boldsymbol{\zeta}_{n,\delta}$ both involve $\{\alpha_\ell\}_{\ell \in [\delta+1]}$ with

$$\alpha_\ell = \mathbb{P}\left(\textbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^\delta, 1; \delta, n-2\right) = \ell - 1\right), \quad \forall \ell \in [\delta+1],$$

where $\{\tilde{\boldsymbol{u}}_i\}_{i=1}^\delta$ are i.i.d. unif(B^{n-2}). α_ℓ is the probability that there are exactly $\ell - 1$ vertices of maximum degree $\delta - 1$ in the random geometric graph $\textbf{Ge}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^\delta, 1; \delta, n-2\right)$.

In principle, α_ℓ can be computed by Monte Carlo method, but the cost increases when δ or n increases. In this section we analytically calculate the α_ℓ for special cases with $\delta = 1$ and $\delta = 2$. For large n and δ, we obtain approximations for α_ℓ and show that CP$(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta})$ is approximately a Poisson.

Example 3.5.1 (Limiting compound Poisson when $\delta = 1$)**.** When $\delta = 1$, $\alpha_2 = 1$ since the number of vertex of the maximum degree 0 is 1. Then $\alpha_1 = 0$ and $\boldsymbol{\zeta}_{n,1}(2) = 1$ and $\boldsymbol{\zeta}_{n,1}(1) = 0$. That is, the increment size of the compound Poisson is always 2. In this case, $\lambda_{n,\delta}(e_{n,\delta}) = \frac{1}{2}\frac{1}{\delta!}(e_{n,\delta})^{\delta}$ and hence the limiting compound Poisson is CP$(\frac{1}{2}\frac{1}{\delta!}(e_{n,\delta})^{\delta}, \delta_{\{2\}})$. That increment size is constant 2 makes sense since $N_{E_1}^{(k)}$, as twice of number of edges, has increment 2 whenever there is a new edge. In terms of $N_{\breve{V}_1}^{(k)}$, the increment is always 2 since the increment always comes with a new pair. The $N_{V_1}^{(k)}$, however, is less obvious. But Theorem 3.2.4 also establishes $N_{V_1}^{(k)}$ has increment 2 in the limit when $p \to \infty$, $\rho \to 1^-$ and $\boldsymbol{\Sigma}$ satisfies some sparsity condition.

As a comparison, Proposition 1 and its proof in [HR11] under row-κ sparsity condition establishes that $N_{E_1}^{(\boldsymbol{R})}/2$ converges to a Pois$(\lambda_{n,\delta}(e_{n,\delta}))$ and $\mathbb{E}N_{V_1}^{(\boldsymbol{R})} \to 2\lambda_{n,\delta}(e_{n,\delta})$ and $\mathbb{P}(N_{V_1}^{(\boldsymbol{R})} > 0) \to 1 - e^{-\lambda_{n,\delta}(e_{n,\delta})}$. Proposition 1 and Proposition 3 in [HR12] under block sparsity condition extend the preceding result to corresponding version in empirical partial correlation graphs, i.e. the same conclusions hold with $\boldsymbol{R}$ replaced by $\boldsymbol{P}$. Our result in Theorem 3.3.11 and Theorem 3.2.4 with $\delta = 1$ characterize the full distribution of the 6 quantities $\{N_i^{(k)} : k \in \{\boldsymbol{R}, \boldsymbol{P}\}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$, and our results in section 3.4 characterizes the first and second moment of them, which together contain the aforementioned previous results. □

Example 3.5.2 (Limiting compound Poisson when $\delta = 2$)**.** When $\delta = 2$, by Lemma 3.5.3, $\alpha_2 = 0$, $\alpha_3 = \frac{3}{2}I_{\frac{3}{4}}(\frac{n-1}{2}, \frac{1}{2})$ and $\alpha_1 = 1 - \alpha_3$, where $I_x(a, b)$ is the regularized incomplete Beta function. Then $\sum_{\ell=1}^{3} \alpha_\ell/\ell = 1 - I_{\frac{3}{4}}(\frac{n-1}{2}, \frac{1}{2})$. Thus the parameters for $CP(\lambda_{n,2}(e_{n,2}), \boldsymbol{\zeta}_{n,2})$ are

$$\boldsymbol{\zeta}_{n,2}(1) = \frac{1 - \frac{3}{2}I_{\frac{3}{4}}(\frac{n-1}{2}, \frac{1}{2})}{1 - I_{\frac{3}{4}}(\frac{n-1}{2}, \frac{1}{2})}, \quad \boldsymbol{\zeta}_{n,2}(2) = 0, \quad \boldsymbol{\zeta}_{n,2}(3) = \frac{\frac{1}{2}I_{\frac{3}{4}}(\frac{n-1}{2}, \frac{1}{2})}{1 - I_{\frac{3}{4}}(\frac{n-1}{2}, \frac{1}{2})} \tag{3.49}$$

and

$$\lambda_{n,2}(e_{n,2}) = \frac{1}{2}(e_{n,2})^2 \left(1 - I_{\frac{3}{4}}\left(\frac{n-1}{2}, \frac{1}{2}\right)\right).$$

Note this corrects an error in Proposition 1 in [HR12], where their incorrect conclusion is built on $N_{E_\delta}^{(\boldsymbol{R})}$ for any $\delta \geq 2$ converges to a Poisson random variable, which is incorrect since we just showed at least when $\delta = 2$, the limit is indeed a compound Poisson $CP(\lambda_{n,2}(e_{n,2}), \boldsymbol{\zeta}_{n,2})$ but not a Poisson.

□

Lemma 3.5.3. *When $\delta = 2$, $\alpha_2 = 0$, $\alpha_3 = \frac{3}{2}I_{\frac{3}{4}}(\frac{n-1}{2}, \frac{1}{2})$ and $\alpha_1 = 1 - \alpha_3$, where $I_x(a, b)$ is the regularized incomplete Beta function.*

In Example 3.5.1 and Example 3.5.2 the limiting compound Poisson for $\delta = 1$ and $\delta = 2$ have been studied. We next show that when n or δ is relatively large, the limiting compound Poisson is approximately a Poisson. For that, the following geometric result is needed.

Lemma 3.5.4. *Let $n \geq 4$ and $\delta \geq 2$.*

(a) Consider $\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta} \overset{i.i.d.}{\sim}$ unif(B^{n-2}). Then

$$\begin{aligned}\mathbb{P}\left(\textbf{\textit{NMD}}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2\right) \geq 1\right) \leq & \delta(n-2)\int_0^1 \left(1-\frac{r^2}{4}\right)^{\frac{(n-2)(\delta-1)}{2}} r^{n-3}dr \\ = & \delta(n-2)2^{n-3}B\left(\frac{1}{4}; \frac{n-2}{2}, \frac{(n-2)(\delta-1)}{2}+1\right),\end{aligned}$$

where $B(\cdot;\cdot,\cdot)$ is the incomplete beta function.

(b)

$$\int_0^1 \left(1-\frac{r^2}{4}\right)^{\frac{(n-2)(\delta-1)}{2}} r^{n-3}dr \leq \begin{cases} \left(\frac{4}{5}\right)^{\frac{(n-2)\delta-1}{2}} + \left(1-\sqrt{\frac{4}{5}}\right)\left(\frac{3}{4}\right)^{\frac{(n-2)(\delta-1)}{2}} & \delta = 2,3, \\ \exp\left(\frac{1}{4}\right)\left(\frac{\delta-1}{\delta}\right)^{\frac{(n-2)(\delta-1)}{2}}\left(\frac{4}{\delta}\right)^{\frac{n-3}{2}} & m \geq 4. \end{cases}$$

Lemma 3.5.4 (a) establishes an upper bound for the probability that there is at least one vertex of maximum degree in the random geometric graph generated by uniform distribution in the unit ball. Lemma 3.5.4 (b) provides a simple upper bound for the integral in (a), and this upper bound provide insight in high dimension (large n) or when there are lots of vertices (large δ). Indeed, when δ is fixed, $\mathbb{P}\left(\textbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2\right) \geq 1\right)$ decays exponentially as n increases as illustrated by Lemma 3.5.4. While n is fixed, it decays as $\delta^{-\frac{n-3}{2}}$ as δ increases.

Recall $\alpha_\ell = \mathbb{P}\left(\textbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2\right) = \ell - 1\right)$ for $\ell \in [\delta+1]$. Then Lemma 3.5.4 immediately yields the following result.

Corollary 3.5.5. *Consider $n \geq 4$ and $\delta \geq 2$.*

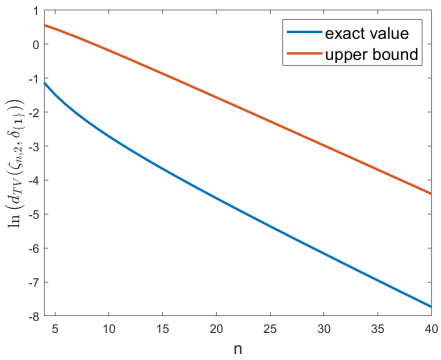
(a)

$$\begin{aligned} d_{TV}\left(\boldsymbol{\zeta}_{n,\delta}, \delta_{\{1\}}\right) \leq \sum_{\ell=2}^{\delta+1}\alpha_\ell \leq & \delta(n-2)2^{n-3}B\left(\frac{1}{4}; \frac{n-2}{2}, \frac{(n-2)(\delta-1)}{2}+1\right) \qquad (3.50) \\ \leq & \begin{cases} \delta(n-2)\left(\left(\frac{4}{5}\right)^{\frac{(n-2)\delta-1}{2}} + \left(1-\sqrt{\frac{4}{5}}\right)\left(\frac{3}{4}\right)^{\frac{(n-2)(\delta-1)}{2}}\right) & \delta = 2,3, \\ \delta(n-2)\exp\left(\frac{1}{4}\right)\left(\frac{\delta-1}{\delta}\right)^{\frac{(n-2)(\delta-1)}{2}}\left(\frac{4}{\delta}\right)^{\frac{n-3}{2}} & \delta \geq 4. \end{cases} \end{aligned}.$$

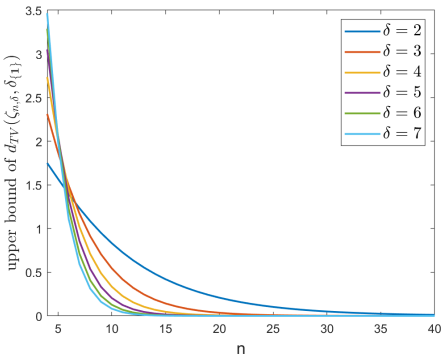
(b) If in addition, $\frac{(e_{n,\delta})^\delta}{\delta!} \leq \gamma_1$, where $e_{n,\delta}$ is as in Theorem 3.2.4, then

$$\left|\lambda_{n,\delta}(e_{n,\delta}) - \frac{(e_{n,\delta})^\delta}{\delta!}\right| \leq \frac{3}{2}\gamma_1 \sum_{\ell=2}^{\delta+1}\alpha_\ell.$$

From Corollary 3.5.5 (a), this total variation distance between the distribution of increment and Dirac measure at 1 decays exponentially as n increases and decays as $\delta^{-\frac{n-3}{2}}$ as δ increases. That means that when either n or δ is large, the limiting compound Poisson is approximately a Poisson. Corollary 3.5.5 (b) states if in addition the threshold is chosen such that $e_{n,\delta}$ satisfies $\frac{(e_{n,\delta})^\delta}{\delta!} \leq \gamma_1$ for any n and δ, then we know the rate of underlying Poisson $\lambda_{n,\delta}(e_{n,\delta})$ is approximately $\frac{1}{\delta!}(e_{n,\delta})^\delta$ with error decaying exponentially as n increases or decaying as $\delta^{-\frac{n-3}{2}}$ as δ increases. In conclusion, $CP(\lambda_{n,\delta}, \boldsymbol{\zeta}_{n,\delta}) \approx \text{Pois}(\frac{(e_{n,\delta})^\delta}{\delta!})$ when n or δ is large and $\frac{(e_{n,\delta})^\delta}{\delta!} \leq \gamma_1$, and in this cases, we do not have to compute the α_ℓ, which involves evaluation of complicated integral.



(a) Log-scale comparison of the decay when $\delta = 2$



(b) Family of upper bounds

Figure 3.4: (a) is a comparison in the log-scale between the upper bound on $d_{TV}(\boldsymbol{\zeta}_{n,2}, \delta_{\{1\}})$ by (3.50) with $\delta = 2$ and the exact value of $d_{TV}(\boldsymbol{\zeta}_{n,2}, \delta_{\{1\}})$ by (3.49). (b) is the plot of the upper bound on $d_{TV}(\boldsymbol{\zeta}_{n,2}, \delta_{\{1\}})$ as a function of n for different values of δ from 2 to 7.

As is shown in the Figure 3.4 (a), for the case $\delta = 2$, $d_{TV}(\boldsymbol{\zeta}_{n,2}, \delta_{\{1\}})$ decays to 0 exponentially and the upper bound in (3.50) also decays to 0 exponentially as n increases. This demonstrate (3.50) captures the rate of decay. Figure 3.4 (b) plots the upper bounds as a function n for fixed δ, and as is clear from the plot, as long as n is above 40, the limiting compound Poisson is approximately a Poisson for any δ. Moreover, as δ increases, the number of samples required for this approximation decreases.

3.5.2 Explicit characterizations for $\alpha(\ell, r_\rho)$

We now turn our attention to the quantity $\alpha(\ell, \rho) = \alpha_{n,\delta}(\ell, \rho)$, which is the parameter of the compound Poisson distribution $\text{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ in Theorem 3.3.11 and the parameter of the second moment in (3.48).

By the discussion following (3.25), we know for $\delta = 1$ and for any n, $\alpha(\ell, r_\rho) = \alpha_{n,1}(\ell, r_\rho) = 1(\delta = 2)$. The next lemma studies the case when $\delta \geq 2$ and is an analogous result to Lemma 3.5.4 and Corollary 3.5.5 (a).

Lemma 3.5.6. *Let $n \geq 4$ and $\delta \geq 2$.*

(a) Consider $\{\boldsymbol{u}'\}_{i=1}^{\delta+1} \overset{i.i.d.}{\sim}$ unif(S^{n-2}). Then for $0 < r < \sqrt{2}$

$$\begin{aligned}
&\mathbb{P}\left(\textbf{\textit{NMD}}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r; \delta+1, n-2\right) \geq 2 | \textit{deg}(\boldsymbol{u}_{\delta+1}') = \delta\right) \\
\leq & \bar{h}\left(\frac{1}{\sqrt{1-r^2/4}}, n, \delta\right) \delta(n-2) \int_0^1 \left(1 - \left(\frac{r_1}{2}\right)^2\right)^{\frac{(n-2)(\delta-1)}{2}} r_1^{n-3} dr_1 \\
= & \bar{h}\left(\frac{1}{\sqrt{1-r^2/4}}, n, \delta\right) \delta(n-2) 2^{n-3} B\left(\frac{1}{4}; \frac{n-2}{2}, \frac{(n-2)(m-1)}{2} + 1\right),
\end{aligned}$$

where $\bar{h}(x, n, \delta) = x^{n+\delta-5} x^{(n-2)(\delta-1)}$.

(b) When $r \leq \begin{cases} 2\sqrt{1-\sqrt{1-1/5}}, & \delta = 2, 3 \\ 2\sqrt{1-\sqrt{1-1/\delta}}, & \delta \geq 4 \end{cases}$,

$$\mathbb{P}\left(\textbf{\textit{NMD}}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r; \delta+1, n-2\right) \geq 2 | \textit{deg}(\boldsymbol{u}_{\delta+1}') = \delta\right) \leq \tilde{h}(n, \delta),$$

where

$$\begin{aligned}
&\tilde{h}(n, \delta) \\
\leq & \begin{cases} \delta(n-2)\{(\sqrt{\frac{5}{4}})^{\frac{\delta-2}{2}} (\sqrt{\frac{4}{5}})^{\frac{(n-2)\delta-1}{2}} + (1 - \sqrt{\frac{4}{5}})(\sqrt{\frac{5}{4}})^{\frac{n+\delta-5}{2}} \left(\frac{3\sqrt{5}}{8}\right)^{\frac{(n-2)(\delta-1)}{2}}\}, & \delta = 2, 3, \\ \delta(n-2)\exp\left(\frac{1}{4}\right) \left(\sqrt{\frac{\delta}{\delta-1}}\right)^{\frac{\delta-2}{2}} \left(\sqrt{\frac{\delta-1}{\delta}}\right)^{\frac{(n-2)(\delta-1)}{2}} \left(\frac{4}{\sqrt{\delta(\delta-1)}}\right)^{\frac{n-3}{2}}, & \delta \geq 4. \end{cases}
\end{aligned}$$

(c) When $\rho \geq 3/5$, $d_{TV}(\boldsymbol{\zeta}_{n,\delta,\rho}, \delta_{\{1\}}) \leq \sum_{\ell=2}^{\delta+1} \alpha(\ell, r_\rho)$, which shares the same upper bounds as in part (a) with r replaced by r_ρ. When $\rho \geq \begin{cases} 4/\sqrt{5} - 1, & \delta = 2, 3 \\ 2\sqrt{1-1/\delta} - 1, & \delta \geq 4 \end{cases}$, $\sum_{\ell=2}^{\delta+1} \alpha(\ell, r_\rho)$ also shares the same upper bound as in part (b).

Lemma 3.5.6 (a) establishes an upper bound for the conditional probability that there is at least two vertices of maximum degree in the random geometric graph generated by uniform distribution on the sphere, conditioned the existence of one such vertex. Lemma 3.5.6 (b) provides a further upper bound when r is relatively small, on the upper bound obtained in part (a), to provide insight in high dimension (large n) or when there are lots of vertices (large δ). Indeed, when δ is fixed, $\mathbb{P}\left(\textbf{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-2\right) \geq 2 | \text{deg}(\boldsymbol{u}_{\delta+1}') = \delta\right)$ decays exponentially as n increases as illustrated by Lemma 3.5.6 (b). While n is fixed, it decays as $\delta^{-\frac{n-3}{2}}$ as δ increases provided that the parameter r decreases in terms of δ as specified in part (b). Lemma 3.5.6 (c) uses the geometric

consequences in part (a) and (b) to establish that the total variation distance between the distribution of increment and Dirac measure at 1 decays exponentially as n increases and decays as $\delta^{-\frac{n-3}{2}}$ as δ increases. That means that when either n or δ is large, and when the threshold ρ is chosen to satisfies the condition in part (c), the compound Poisson approximation in Theorem 3.3.11 is approximately a Poisson.

By examining the proof, Lemma 3.5.6 (c) essentially proves that $\alpha_\ell(1, r_\rho) \approx 1$ and

$$\sum_{\ell=2}^{\delta+1} \alpha_\ell(1, r_\rho) \approx 0.$$

In this case it is not difficult to see that the $\lambda_{p,n,\delta,\rho}$ satisfies

$$\lambda_{p,n,\delta,\rho} \approx \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta,$$

which then implies $\mathrm{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho}) \approx \mathrm{Pois}(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta)$. Moreover the second moment of $\mathrm{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ in (3.48) approximately equals to

$$\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta + \left(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\right)^2.$$

One can obtain the errors of the approximations in the preceding two displays analogous to Lemma 3.5.6 (c), and these straightforward extensions are omitted.

We have shown that the limiting compound Poisson $\mathrm{CP}(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta})$ in Theorem 3.2.4 can be approximately by $\mathrm{Pois}(\frac{(e_{n,\delta})^\delta}{\delta!})$ and the non-asymptotic compound Poisson $\mathrm{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ in Theorem 3.3.11 can be approximated by $\mathrm{Pois}(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta)$ with small errors for large n or δ. Figure 3.9 in is a numerical simulation to demonstrate the effect of using Poisson distributions to approximate the distributions of random quantities in $\{N_i^{(k)} : k = \boldsymbol{R}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$.

3.6 Conclusions and discussions

In this chapter, we studied the number of hubs in both the empirical correlation graph and the empirical partial correlation graph in a unified framework. To be specific, we show the number of hubs in terms of $N_{V_\delta}^{(k)}$ or $N_{\breve{V}_\delta}^{(k)}$ and the star subgraph counts $N_{E_\delta}^{(k)}$ both are close to a common compound Poisson in distribution, asymptotically and non-asymptotically. We also establish that the first and second moments of random quantities of interest are close to that of the compound Poisson. The parameters in the compound Poisson are characterized in closed form in terms of quantities from a random geometric graph. The parameters are also approximated by simple formulae, which implies the compound Poisson can be approximated by a Poisson for reasonably large sample size

n or reasonably hub degree δ.

In Subsection 3.3.2 we represent the empirical correlation graph as random pseudo geometric graph with $\boldsymbol{U}$-scores $\{\boldsymbol{u}_i\}_{i=1}^p$ as vertices. Then $N_{E_\delta}^{(\boldsymbol{R})}$ is the star subgraph counts and $N_{V_\delta}^{(\boldsymbol{R})}$ is the number of vertices of degree at least δ of the random pseudo geometric graph. Note the monograph [Pen03] studied the number of *induced* subgraphs isomorphic to a given graph, typical vertices, and other graphical quantities of random geometric graph thoroughly, where they assume the vertices of the random geometric graph are i.i.d. on $\mathbb{R}^d$. It is worth pointing out that in the Example after Corollary 3.6 in [Pen03], the author comments on the number of vertices of degree at least 3 is asymptotically a compound Poisson without characterizing the parameters of the compound Poisson. Here in Theorem 3.2.4 and Theorem 3.3.11, we give characterizations of the compound Poisson distributions with closed-form formulae. Moreover, it is clear that our random pseudo geometric graph model is different from the classic random geometric graph since our vertices $\boldsymbol{u}_i$ lie in the unit sphere instead of the whole Euclidean space and our distance is $\mathbf{dist}(\cdot,\cdot)$ instead of Euclidean distance. We also want to emphasize a key difference is that no independence among all vertices $\boldsymbol{u}_i$ are imposed in our model. Indeed in our model, the correlations between vertices $\boldsymbol{u}_i$ are encoded by a sparse $\boldsymbol{\Sigma}$.

Future directions include generalizing the results to non-sparse $\boldsymbol{\Sigma}$ since it is observed from simulations that the same compound Poisson characterization also holds for $\boldsymbol{\Sigma}$ dense but with many entries of small magnitude. Another is to extend the convergence of the first and second moments to higher moments as already discussed at the end of Section 3.4. A third direction is to develop potential applications in the hypothesis testing to test whether the dispersion matrix $\boldsymbol{\Sigma}$ satisfies a certain structure based on the compound Poisson characterizations.

3.7 Controlling local normalized determinant by extreme eigenvalues

Lemma 3.7.1. *(a) For any symmetric positive definite $\boldsymbol{A}$, $\mu_{n,m}(\boldsymbol{A})$ is bounded by powers of the largest local condition number:*

$$\mu_{n,m}(\boldsymbol{A}) \leq \begin{cases} \max\limits_{\mathcal{I}\subset[p]} \left(\frac{\lambda_{max}(\boldsymbol{A}_{\mathcal{I}})}{\lambda_{min}(\boldsymbol{A}_{\mathcal{I}})}\right)^{\frac{m(n-1)}{2}}, & \boldsymbol{A} \text{ not diagonal,} \\ 1, & \boldsymbol{A} \text{diagonal.} \end{cases}$$

(b) For any symmetric positive definite $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, $\mu_{n,m}(\boldsymbol{A})$ is bounded by powers of the condition number:

$$\mu_{n,m}(\boldsymbol{A}) \leq \begin{cases} \left(\frac{\lambda_{max}(\boldsymbol{A})}{\lambda_{min}(\boldsymbol{A})}\right)^{\frac{m(n-1)}{2}}, & \boldsymbol{A} \text{ not diagonal,} \\ 1, & \boldsymbol{A} \text{ diagonal.} \end{cases}$$

(c) Consider a sequence of symmetric positive definite matrices $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ with increasing dimension p. If $\lambda_{min}(\mathbf{\Sigma}) \geq \underline{\lambda}$ and $\lambda_{max}(\mathbf{\Sigma}) \leq \overline{\lambda}$ for all p, then

$$\mu_{n,m}(\mathbf{\Sigma}) \leq \begin{cases} \left(\frac{\overline{\lambda}}{\underline{\lambda}}\right)^{\frac{m(n-1)}{2}}, & \mathbf{\Sigma} \text{ not diagonal,} \\ 1, & \mathbf{\Sigma} \text{ diagonal.} \end{cases}$$

(d) Consider a sequence of symmetric positive definite matrices $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ with increasing dimension p. Let $M > 0$ be a constant. Suppose the variance of each variable is uniformly bounded by M, i.e. for all p, $\sup_{1\leq i\leq p}(\mathbf{\Sigma})_{ii} \leq M$. Moreover suppose $\lambda_{min}(\mathbf{\Sigma}) \geq \underline{\lambda}$ for all p. Then

$$\mu_{n,m}(\mathbf{\Sigma}) \leq \begin{cases} \left(\frac{Mm}{\underline{\lambda}}\right)^{\frac{m(n-1)}{2}}, & \mathbf{\Sigma} \text{ not diagonal,} \\ 1, & \mathbf{\Sigma} \text{ diagonal.} \end{cases}$$

Proof: (a) It follows directly by definition of $\mu_{n,m}(A)$ and $\mu(A_{\mathcal{I}}) \geq \left(\frac{\lambda_{\min}(A_{\mathcal{I}})}{\lambda_{\max}(A_{\mathcal{I}})}\right)^m$.

(b) Since $\mu_{n,m}(A)$ is inceasing in m as discussed after Example 3.2.8, $\mu_{n,m}(A) \leq \mu_{n,p}(A)$. The proof is then complete by applying (a) to $\mu_{n,p}(A)$.

(c) It follows directly from (b).

(d) Let $\mathcal{I} \subset [p]$ with $|\mathcal{I}| = m$. Since $\boldsymbol{A}_{\mathcal{I}}$ is symmetric positive definite,

$$|(\boldsymbol{A}_{\mathcal{I}})_{ij}| \leq \sqrt{(\boldsymbol{A}_{\mathcal{I}})_{ii}(\boldsymbol{A}_{\mathcal{I}})_{jj}} \leq M$$

and thus $\lambda_{\max}(\boldsymbol{A}_{\mathcal{I}}) = \|\boldsymbol{A}_{\mathcal{I}}\|_2 \leq \|\boldsymbol{A}_{\mathcal{I}}\|_F \leq Mm$. The remaining of the proof is similar to that of (b).

□

3.8 Proofs in Subsection 3.3.1 and Subsection 3.3.2

3.8.1 Proof of Lemma 3.3.2

Proof of Lemma 3.3.2: $\tilde{\boldsymbol{X}}$, defined in (3.15), is of rank $n-1$ with probability 1, since it has a density with respect to Lebesgue measure on $\mathbb{R}^{(n-1)p}$ and $p \geq n$. Then $\boldsymbol{U}$ have rank $n-1$ with probability 1 as well since $\boldsymbol{U}$ is obtained by normalizing the columns of $\tilde{\boldsymbol{X}}$. Thus $\boldsymbol{B}$ is of the rank $n-1$ with probability 1. □

3.8.2 Proof of Lemma 3.3.3 (b)

It suffices to prove the following statement:

$$f_{\boldsymbol{u}_{j_1},\boldsymbol{u}_{j_2},\cdots,\boldsymbol{u}_{j_m}}(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m) \le \mu_{n,m}(\boldsymbol{\Sigma}_{\mathcal{J}}), \quad \forall\, \boldsymbol{v}_i \in S^{n-2}, \forall\, i \in [m].$$

For notation convenience, we only present the proof for $m = p$ and $\mathcal{J} = [p]$ since the proof of the general m and $\mathcal{J}$ follows the same proof procedure. When $m = p$ and $\mathcal{J} = [p]$, the statement of Lemma 3.3.3 (b) become:

The joint density of columns of $\boldsymbol{U}$-score w.r.t. $\otimes^p \sigma^{n-1}$ is upper bounded by $\mu_{n,p}(\boldsymbol{\Sigma})$:

$$f_{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_p}(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_p) \le \mu_{n,p}(\boldsymbol{\Sigma}), \quad \forall\, \boldsymbol{v}_i \in S^{n-2}, \forall\, i \in [p]. \tag{3.51}$$

Proof of (3.51)**:** Recall $\{\tilde{\boldsymbol{x}}^{(i)}\}_{i=1}^{n-1} \subset \mathbb{R}^p$, the rows of $\tilde{\boldsymbol{X}}$, are i.i.d. copy of $\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$. Let $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^p$ be the columns of $\tilde{\boldsymbol{X}}$. Then $\boldsymbol{u}_i := \frac{\tilde{\boldsymbol{x}}_i}{\|\tilde{\boldsymbol{x}}_i\|_2} \in \mathbb{R}^{n-1}$ has distribution unif(S^{n-2}) for $i \in [p]$.

When $\boldsymbol{\Sigma}$ is symmetric positive definite and diagonal, $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^p$ are independent, which imply $\{\boldsymbol{u}_i\}_{i=1}^p$ are independent. Thus in this case, the joint density of columns of $\boldsymbol{U}$-score w.r.t. $\otimes^p \sigma^{n-1}$ is 1.

Consider general symmetric positive definite $\boldsymbol{\Sigma}$. The probability density of $\tilde{\boldsymbol{X}}$ w.r.t. the Lebesgue measure on $\mathbb{R}^{(n-1)p}$ is

$$f_{\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{X}}) = \det(\boldsymbol{\Sigma})^{-\frac{n-1}{2}}(2\pi)^{-\frac{(n-1)p}{2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n-1}\left(\tilde{\boldsymbol{x}}^{(j)}\right)^T \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}^{(j)}\right).$$

Use the spherical transform for each column $\tilde{\boldsymbol{x}}_i = \left(\tilde{\boldsymbol{X}}_{ji} : 1 \le j \le n-1\right)^T$:

$$\begin{cases} \tilde{\boldsymbol{X}}_{1i} = R_i \cos(\theta_{1i}), \\ \tilde{\boldsymbol{X}}_{2i} = R_i \sin(\theta_{1i})\cos(\theta_{2i}), \\ \vdots \\ \tilde{\boldsymbol{X}}_{(n-2)i} = R_i \sin(\theta_{1i})\sin(\theta_{2i})\cdots\sin(\theta_{(n-3)i})\cos(\theta_{(n-2)i}), \\ \tilde{\boldsymbol{X}}_{(n-1)i} = R_i \sin(\theta_{1i})\sin(\theta_{2i})\cdots\sin(\theta_{(n-3)i})\sin(\theta_{(n-2)i}), \end{cases} \quad \text{for } 1 \le i \le p,$$

where for each $i \in [p]$: $R_i \ge 0, \theta_{ji} \in [0,\pi]$ for $1 \le j \le n-3$ and $\theta_{(n-2)i} \in [0, 2\pi)$.

Denote $\boldsymbol{R} = (R_i : 1 \le i \le p)$ and $\boldsymbol{\Theta} = (\theta_{ji} : 1 \le i \le p, 1 \le j \le (n-2))$. Then the joint density of $(\boldsymbol{R},\boldsymbol{\Theta})$ is:

$$f_{\boldsymbol{R},\boldsymbol{\Theta}}(\boldsymbol{R},\boldsymbol{\Theta})$$

$$=\det(\boldsymbol{\Sigma})^{-\frac{n-1}{2}}(2\pi)^{-\frac{(n-1)p}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{n-1}\left(\boldsymbol{h}^{(j)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{h}^{(j)}\right)\prod_{i=1}^{p}\left(R_i^{n-2}\prod_{j=1}^{n-2}\sin^{n-2-j}(\theta_{ji})\right),$$

where

$$\boldsymbol{h}^{(j)}=\left(R_i\cos(\theta_{ji})\prod_{q=1}^{j-1}\sin(\theta_{qi}):1\le i\le p\right)^T\in\mathbb{R}^p \text{ for } 1\le j\le n-2$$

and

$$\boldsymbol{h}^{(n-1)}=\left(R_i\prod_{q=1}^{n-2}\sin(\theta_{qi}):1\le i\le p\right)^T\in\mathbb{R}^p.$$

Then the density of $\boldsymbol{\Theta}$ is:

$$\begin{aligned}
&f_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})\\
=&\det(\boldsymbol{\Sigma})^{-\frac{n-1}{2}}(2\pi)^{-\frac{(n-1)p}{2}}\prod_{i=1}^{p}\prod_{j=1}^{n-2}\sin^{n-2-j}(\theta_{ji})\int_{[0,\infty)^p}e^{-\frac{1}{2}\sum_{j=1}^{n-1}\left(\boldsymbol{h}^{(j)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{h}^{(j)}}\prod_{i=1}^{p}\left(R_i^{n-2}\right)\prod_{i=1}^{p}dR_i\\
\le&\det(\boldsymbol{\Sigma})^{-\frac{n-1}{2}}(2\pi)^{-\frac{(n-1)p}{2}}\prod_{i=1}^{p}\prod_{j=1}^{n-2}\sin^{n-2-j}(\theta_{ji})\int_{[0,\infty)^p}e^{-\frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\sum_{j=1}^{n-1}\|\boldsymbol{h}^{(j)}\|_2^2}\prod_{i=1}^{p}\left(R_i^{n-2}\right)\prod_{i=1}^{p}dR_i\\
=&\det(\boldsymbol{\Sigma})^{-\frac{n-1}{2}}(2\pi)^{-\frac{(n-1)p}{2}}\prod_{i=1}^{p}\prod_{j=1}^{n-2}\sin^{n-2-j}(\theta_{ji})\int_{[0,\infty)^p}e^{-\frac{1}{2}[\lambda_{\max}(\boldsymbol{\Sigma})]^{-1}\sum_{i=1}^{p}R_i^2}\prod_{i=1}^{p}\left(R_i^{n-2}\right)\prod_{i=1}^{p}dR_i\\
=&\det(\boldsymbol{\Sigma})^{-\frac{n-1}{2}}(2\pi)^{-\frac{(n-1)p}{2}}\prod_{i=1}^{p}\prod_{j=1}^{n-2}\sin^{n-2-j}(\theta_{ji})\left(\int_{[0,\infty)}e^{-\frac{1}{2}[\lambda_{\max}(\boldsymbol{\Sigma})]^{-1}R_1^2}R_1^{n-2}dR_1\right)^p\\
\overset{(m)}{=}&\det(\boldsymbol{\Sigma})^{-\frac{n-1}{2}}(2\pi)^{-\frac{(n-1)p}{2}}\prod_{i=1}^{p}\prod_{j=1}^{n-2}\sin^{n-2-j}(\theta_{ji})\left([\lambda_{\max}(\boldsymbol{\Sigma})]^{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-3}{2}}\right)^p\\
\overset{(mm)}{=}&\left[\frac{(\lambda_{\max}(\boldsymbol{\Sigma}))^p}{\det(\boldsymbol{\Sigma})}\right]^{\frac{n-1}{2}}\frac{1}{|S^{n-2}|^p}\prod_{i=1}^{p}\prod_{j=1}^{n-2}\sin^{n-2-j}(\theta_{ji})
\end{aligned}$$

where equality (m) follows from the integration of Chi distribution with degree $n-1$, and equality (mm) follows from $|S^{(n-2)}|=2\pi^{\frac{n-1}{2}}/\Gamma((n-1)/2)$. The proof is complete by noticing $f_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$ is joint density of columns of $\boldsymbol{U}$-score expressed in spherical coordinate and

$$\frac{1}{|S^{n-2}|^p}\prod_{i=1}^{p}\prod_{j=1}^{n-2}\sin^{n-2-j}(\theta_{ji})$$

is the joint distribution of p independent unif(S^{n-2}) expressed in spherical coordinate. □

3.9 Proof of Proposition 3.3.6

3.9.1 Auxiliary lemmas for Proposition 3.3.6

Recall for any $\delta \geq 1$, $C_\delta^<$ is defined in (3.29). For $\vec{i} \in C_\ell^<$, define a symmetric positive definite matrix $\mathbf{\Sigma}_{\vec{i}} \in \mathbb{R}^{(\ell+1)\times(\ell+1)}$ to be the submatrix of $\mathbf{\Sigma}$, consisting of rows and columns $\mathbf{\Sigma}$ indexed by the ordered components $(i_0, i_1, \ldots, i_\ell)$ of $\vec{i}$. Let $\left[\vec{i}\right] = \{i_0, i_1, \cdots, i_\ell\}$ be the unordered set of indexes of any $\vec{i} \in C_\ell^<$. Then $\mathbf{\Sigma}_{\vec{i}} \in \mathbf{\Sigma}_{[\vec{i}]}$ and $\mu_{n,\ell+1}\left(\mathbf{\Sigma}_{\vec{i}}\right) = \mu_{n,\ell+1}\left(\mathbf{\Sigma}_{[\vec{i}]}\right)$, where $\mathbf{\Sigma}_{[\vec{i}]}$ and $\mu_{n,\ell+1}\left(\mathbf{\Sigma}_{[\vec{i}]}\right)$ are defined in the paragraph after Definition 3.2.3.

Lemma 3.9.1. *Suppose $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \mathbf{\Sigma}, \theta)$. Let $\ell \in [p-1]$. Consider $\vec{i} = (i_0, i_1, \cdots, i_\ell) \in C_\ell^<$.*

$$\mathbb{E}\prod_{q=1}^{\ell} \Phi_{i_0 i_q}^{(\boldsymbol{R})} \leq \mu_{n,\ell+1}\left(\mathbf{\Sigma}_{\vec{i}}\right)\left(2P_n(r_\rho)\right)^\ell.$$

Moreover, in last display the equality holds and $\mu_{n,\ell+1}\left(\mathbf{\Sigma}_{\vec{i}}\right) = 1$ when $\mathbf{\Sigma}_{\vec{i}}$ is diagonal.

Proof: $\prod_{q=1}^{\ell} \Phi_{i_0 i_q}^{(\boldsymbol{R})}$ is a nonnegative Borel Measurable function of $\boldsymbol{u}_j$ for $j \in \left[\vec{i}\right]$. By Lemma 3.3.3 (c), it suffices to show

$$\mathbb{E}\prod_{q=1}^{\ell} \Phi_{i_0 i_q}^{(\boldsymbol{R})} \leq \left(2P_n(r)\right)^\ell$$

for the case $\boldsymbol{u}_j$ for $j \in \left[\vec{i}\right]$ are $\ell+1$ independent unif(S^{n-2}). The last inequality indeed holds with equality, which follows from that the terms in the product on the left side are independent conditioned on $\boldsymbol{u}_i$. □

Lemma 3.9.1 suggests differentiating whether $\mathbf{\Sigma}_{\vec{i}}$ is diagonal or not since $\mu_{n,\ell+1}\left(\mathbf{\Sigma}_{\vec{i}}\right) = 1$ when $\mathbf{\Sigma}_{\vec{i}}$ is diagonal. The next lemma is to establish in the worser case when $\mathbf{\Sigma}_{\vec{i}}$ is not diagonal, the number of such terms are not too many.

Lemma 3.9.2. *Let $\mathbf{\Sigma}$ be row-κ sparse. Let $\delta \in [p-1]$. Then*

$$\sum_{\substack{\vec{i} \in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1 \leq \frac{\delta(\delta+1)}{2}(\kappa-1)\binom{p}{\delta} \leq \frac{(\delta+1)}{2((\delta-1)!)}p^\delta(\kappa-1).$$

Proof: Note that

$$\sum_{\substack{\vec{i} \in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}} 1 \geq \frac{1}{\delta!}p(p-\kappa)\ldots(p-\delta\kappa),$$

where the $\frac{1}{\delta!}$ is due to in our definition $\vec{i}$ the index $i_1 < \ldots < i_\delta$ are sorted. Then

$$\sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1 \le \binom{p}{1}\binom{p-1}{\delta} - \frac{1}{\delta!}p\prod_{\ell=1}^{\delta}(p-\ell\kappa) \le \frac{\delta(\delta+1)}{2}(\kappa-1)\binom{p}{\delta},$$

where the last inequality follows from Lemma 3.14.3 (b). □

Note $\kappa = 1$, the Lemma 3.9.2 shows $\sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1 = 0$, which means $\mathbf{\Sigma}$ is diagonal matrix. Next we present a lemma to bound $\sum_{\vec{i}\in C_\ell^<} \mu_{n,\ell+1}\left(\mathbf{\Sigma}_{\vec{i}}\right)$.

Lemma 3.9.3.

$$\sum_{\vec{i}\in C_\ell^<} \mu_{n,\ell+1}\left(\mathbf{\Sigma}_{\vec{i}}\right) \le \frac{p^{\ell+1}}{\ell!}\left(1+\ell^2\mu_{n,\ell+1}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right).$$

Proof:

$$\begin{aligned}\sum_{\vec{i}\in C_\ell^<} \mu_{n,\ell+1}(\mathbf{\Sigma}_{\vec{i}}) =& \sum_{\substack{\vec{i}\in C_\ell^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}} 1 + \sum_{\substack{\vec{i}\in C_\ell^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} \mu_{n,\ell+1}(\mathbf{\Sigma}_{\vec{i}}) \\ \le& \binom{p}{1}\binom{p-1}{\ell} + \mu_{n,\ell+1}(\mathbf{\Sigma}) \sum_{\substack{\vec{i}\in C_\ell^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1 \\ \le& \frac{p^{\ell+1}}{\ell!}\left(1+\ell^2\mu_{n,\ell+1}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right),\end{aligned}$$

where the first inequality follows from the Lemma 3.3.3 (a), and the second inequality follows from Lemma 3.9.2. □

Lemma 3.9.4. *Let $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \mathbf{\Sigma}, \theta)$. Let $\{i_q\}_{q=0}^{\alpha}, \{j_q\}_{q=0}^{\beta} \subset [p]$ be respectively a sequence of $\alpha+1$ and $\beta+1$ distinct integers. Let $m \in [\min\{\alpha,\beta\}]$. Suppose $i_q = j_q$ for $q \in [m]$ and $i_q \neq j_{q'}$ for $q, q' \notin [m]$. Denote $\mathcal{I} = \bigcup_{q=0}^{\alpha}\{i_q\} \bigcup \left(\bigcup_{q'=0}^{\beta}\{j_q'\}\right)$ and then $|\mathcal{I}| = \alpha+\beta-m+2$.*

(a) Then

$$\mathbb{E}\left(\prod_{q=1}^{\alpha}\Phi_{i_0 i_q}^{(\boldsymbol{R})}\right)\left(\prod_{q'=1}^{\beta}\Phi_{j_0 j_{q'}}^{(\boldsymbol{R})}\right) \le \mu_{n,|\mathcal{I}|}\left(\mathbf{\Sigma}_{\mathcal{I}}\right)\left(2P_n(r_\rho)\right)^{\alpha+\beta-m}\left(2P_n(2r_\rho)\right). \tag{3.52}$$

(b) Then

$$\mathbb{E}\Phi_{i_0 j_0}^{(\boldsymbol{R})}\left(\prod_{q=1}^{\alpha}\Phi_{i_0 i_q}^{(\boldsymbol{R})}\right)\left(\prod_{q'=1}^{\beta}\Phi_{j_0 j_{q'}}^{(\boldsymbol{R})}\right) \le \mu_{n,|\mathcal{I}|}\left(\mathbf{\Sigma}_{\mathcal{I}}\right)\left(2P_n(r_\rho)\right)^{\alpha+\beta-m+1}. \tag{3.53}$$

(3.53) *also holds with* $m = 0$.

Proof: (a) By Lemma 3.3.3 (c), it suffices to prove (3.52) without $\mu_{n,|\mathcal{I}|}(\boldsymbol{\Sigma}_{\mathcal{I}})$ for the case $\{\boldsymbol{u}_j\}$ for $j \in \mathcal{I}$ are independent unif(S^{n-2}). Conditioned on $\boldsymbol{u}_{i_0}$ and $\boldsymbol{u}_{j_0}$, $\{\Phi^{(\boldsymbol{R})}_{i_0 i_q}\Phi^{(\boldsymbol{R})}_{j_0 i_q}\}_{q=1}^{m}$ are $i.i.d.$, $\{\Phi^{(\boldsymbol{R})}_{i_0 i_q}\}_{q=m+1}^{\alpha} \bigcup \{\Phi^{(\boldsymbol{R})}_{j_0 j_{q'}}\}_{q'=m+1}^{\beta}$ are $i.i.d.$ and moreover, every term in $\{\Phi^{(\boldsymbol{R})}_{i_0 i_q}\Phi^{(\boldsymbol{R})}_{j_0 i_q}\}_{q=1}^{m}$ is independent of every term in $\{\Phi^{(\boldsymbol{R})}_{i_0 i_q}\}_{q=m+1}^{\alpha} \bigcup \{\Phi^{(\boldsymbol{R})}_{j_0 j_{q'}}\}_{q'=m+1}^{\beta}$. Thus

$$
\begin{aligned}
&\mathbb{E}\left[\left(\prod_{q=1}^{\alpha}\Phi^{(\boldsymbol{R})}_{i_0 i_q}\right)\left(\prod_{q'=1}^{\beta}\Phi^{(\boldsymbol{R})}_{j_0 j_{q'}}\right)\middle|\, \boldsymbol{u}_{i_0}, \boldsymbol{u}_{j_0}\right] \\
=&(\mathbb{E}[\Phi^{(\boldsymbol{R})}_{i_0 i_1}\Phi^{(\boldsymbol{R})}_{j_0 i_1}\Big|\, \boldsymbol{u}_{i_0}, \boldsymbol{u}_{j_0}])^{m}(\mathbb{E}[\Phi^{(\boldsymbol{R})}_{i_0 i_\alpha}1(\alpha > m) + \Phi^{(\boldsymbol{R})}_{j_0 j_\beta}1(\alpha = m, \beta > m)\Big|\, \boldsymbol{u}_{i_0}, \boldsymbol{u}_{j_0}])^{\alpha+\beta-2m} \\
=&\left(\mathbb{E}\left[\Phi^{(\boldsymbol{R})}_{i_0 i_1}\Phi^{(\boldsymbol{R})}_{j_0 i_1}\middle|\, \boldsymbol{u}_{i_0}, \boldsymbol{u}_{j_0}\right]\right)^{m}(2P_n(r_\rho))^{\alpha+\beta-2m}, \qquad (3.54)
\end{aligned}
$$

where for the first equality the convention $0^0 = 1$ is used if $\alpha = \beta = m$. Notice (3.54) also holds for $m = 0$.

Denote $\overline{\mathbf{SC}}(r, \boldsymbol{z}) = \mathbf{SC}(r, \boldsymbol{z}) \cup \mathbf{SC}(r, -\boldsymbol{z})$. Then conditioned on $\boldsymbol{u}_{i_0}$ and $\boldsymbol{u}_{j_0}$,

$$
\begin{aligned}
&\Phi^{(\boldsymbol{R})}_{i_0 i_1}\Phi^{(\boldsymbol{R})}_{j_0 i_1} \\
=&1\left(\boldsymbol{u}_{i_1} \in \overline{\mathbf{SC}}(r, \boldsymbol{u}_{i_0}) \cap \overline{\mathbf{SC}}(r, \boldsymbol{u}_{j_0})\right) \\
=&1(\|\boldsymbol{u}_{i_0} - \boldsymbol{u}_{j_0}\|_2 \le 2r_\rho \text{ or } \|\boldsymbol{u}_{i_0} + \boldsymbol{u}_{j_0}\|_2 \le 2r_\rho)1\left(\boldsymbol{u}_{i_1} \in \overline{\mathbf{SC}}(r, \boldsymbol{u}_{i_0}) \cap \overline{\mathbf{SC}}(r, \boldsymbol{u}_{j_0})\right),
\end{aligned}
$$

where the last equality follows from $\overline{\mathbf{SC}}(r_\rho, \boldsymbol{u}_{i_0}) \cap \overline{\mathbf{SC}}(r_\rho, \boldsymbol{u}_{j_0})$ is non-empty only when $\|\boldsymbol{u}_{i_0} - \boldsymbol{u}_{j_0}\|_2 \le 2r_\rho$ or $\|\boldsymbol{u}_{i_0} + \boldsymbol{u}_{j_0}\|_2 \le 2r_\rho$. Plug the above inequality into (3.54),

$$
\begin{aligned}
&\mathbb{E}\left[\left(\prod_{q=1}^{\alpha}\Phi^{(\boldsymbol{R})}_{i_0 i_q}\right)\left(\prod_{q'=1}^{\beta}\Phi^{(\boldsymbol{R})}_{j_0 j_{q'}}\right)\middle|\, \boldsymbol{u}_{i_0}, \boldsymbol{u}_{j_0}\right] \\
=&1(\|\boldsymbol{u}_{i_0} - \boldsymbol{u}_{j_0}\|_2 \le 2r_\rho \text{ or } \|\boldsymbol{u}_{i_0} + \boldsymbol{u}_{j_0}\|_2 \le 2r_\rho)(\mathbb{E}[\Phi^{(\boldsymbol{R})}_{i_0 i_1}\Phi^{(\boldsymbol{R})}_{j_0 i_1}\Big|\, \boldsymbol{u}_{i_0}, \boldsymbol{u}_{j_0}])^{m}(2P_n(r_\rho))^{\alpha+\beta-2m} \\
\le&1(\|\boldsymbol{u}_{i_0} - \boldsymbol{u}_{j_0}\|_2 \le 2r_\rho \text{ or } \|\boldsymbol{u}_{i_0} + \boldsymbol{u}_{j_0}\|_2 \le 2r_\rho)(2P_n(r_\rho))^{m}(2P_n(r_\rho))^{\alpha+\beta-2m}.
\end{aligned}
$$

The result then follows by taking expectation w.r.t. $\boldsymbol{u}_{i_0}$ and $\boldsymbol{u}_{j_0}$.

(b) Similar to proof of (a), it suffices to prove (3.53) without $\mu_{n,|\mathcal{I}|}(\boldsymbol{\Sigma}_{\mathcal{I}})$ for the case $\boldsymbol{u}_j$ for $j \in \mathcal{I}$ are independent unif(S^{n-2}). Conditioned on $\boldsymbol{u}_{i_0}$ and $\boldsymbol{u}_{j_0}$,

$$
\mathbb{E}\left[\Phi^{(\boldsymbol{R})}_{i_0 j_0}\left(\prod_{q=1}^{\alpha}\Phi^{(\boldsymbol{R})}_{i_0 i_q}\right)\left(\prod_{q'=1}^{\beta}\Phi^{(\boldsymbol{R})}_{j_0 j_{q'}}\right)\middle|\, \boldsymbol{u}_{i_0}, \boldsymbol{u}_{j_0}\right]
$$

$$
\begin{aligned}
=&\Phi^{(\boldsymbol{R})}_{i_0 j_0}\left(\mathbb{E}\left[\Phi^{(\boldsymbol{R})}_{i_0 i_1}\Phi^{(\boldsymbol{R})}_{j_0 i_1}\middle|\boldsymbol{u}_{i_0},\boldsymbol{u}_{j_0}\right]\right)^m (2P_n(r_\rho))^{\alpha+\beta-2m}\\
\leq&\Phi^{(\boldsymbol{R})}_{i_0 j_0}(2P_n(r_\rho))^m (2P_n(r_\rho))^{\alpha+\beta-2m},
\end{aligned}
$$

where the equality follows from (3.54). The result then follows by taking expectation w.r.t. $\boldsymbol{u}_{i_0}$ and $\boldsymbol{u}_{j_0}$. Notice (3.54) also holds for $m=0$.

□

3.9.2 Lemmas on double summations

Denote $\vec{i}\cup\vec{j}=\left[\vec{i}\right]\bigcup\left[\vec{j}\right]$ for any $\vec{i}\in C_q^<$ and any $\vec{j}\in C_\delta^<$. Consider any $\theta_{\vec{i},\vec{j}}$ that is a non-negative function of $\boldsymbol{u}_\ell$ for $\ell\in\vec{i}\cup\vec{j}$ defined for $\vec{i}\in C_q^<$ and $\vec{j}\in C_\delta^<$ with $1\leq\delta\leq q\leq p-1$. In this section an upper bound on $\mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in C_\delta^<}\theta_{\vec{i},\vec{j}}$ is presented. The results in this subsection will be used in the proofs of Proposition 3.3.6 and Proposition 3.4.3.

For $i\in[p]$, let

$$
\mathcal{NZ}(i):=\{m\in[p]:\boldsymbol{\Sigma}_{im}\neq 0\} \tag{3.55}
$$

denote the index of the variables that has non zero correlation with the i-th variable. For $\vec{i}\in C_q^<$, define $\mathcal{NZ}\left(\vec{i}\right):=\bigcup_{\ell=0}^{q}\mathcal{NZ}(i_\ell)$. Since $\boldsymbol{\Sigma}$ is row-κ sparse, for any $\vec{i}\in C_q^<$, $\left|\mathcal{NZ}\left(\vec{i}\right)\right|\leq(q+1)\kappa$, and

$$
p_{\vec{i}}:=\left|[p]\backslash\mathcal{NZ}\left(\vec{i}\right)\right|\geq p-(q+1)\kappa. \tag{3.56}
$$

Note that $p_{\vec{i}}$ is the number of variables that are independent of variables in the group $[\vec{i}]$.

For $\vec{i}\in C_q^<$, define

$$
J_{\vec{i}}:=\left\{\vec{j}\in C_\delta^<:\bigcup_{\ell=0}^{\delta}\{j_\ell\}\subset\bigcup_{\ell=0}^{q}\{i_\ell\}\right\}, \tag{3.57}
$$

$$
T_{\vec{i}}:=\left\{\vec{j}\in C_\delta^<:\left(\bigcup_{\ell=0}^{\delta}\{j_\ell\}\right)\bigcap\left(\mathcal{NZ}\left(\vec{i}\right)\right)=\emptyset\right\}, \tag{3.58}
$$

$$
N_{\vec{i}}:=C_q^<\backslash J_{\vec{i}}\backslash T_{\vec{i}}. \tag{3.59}
$$

Here $J_{\vec{i}}$ is the set of indexes in $C_\delta^<$ consisting of coordinates as subsets of $\left[\vec{i}\right]$; $T_{\vec{i}}$ is the set of indexes in $C_\delta^<$ consisting of coordinates outside neighborhood of $\vec{i}$; $N_{\vec{i}}$ is the set of "correlated but not highly correlated" indexes in $C_\delta^<$, i.e. the set of indexes of which at least one coordinate is in the neighborhood of $\vec{i}$, but excluding those sets of indexes of which the set of coordinates are subsets as that of $\vec{i}$.

The strategy is to decompose

$$\mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in C_\delta^<}\theta_{\vec{i},\vec{j}} = \mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in J_{\vec{i}}}\theta_{\vec{i},\vec{j}} + \mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in T_{\vec{i}}}\theta_{\vec{i},\vec{j}} + \mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in N_{\vec{i}}}\theta_{\vec{i},\vec{j}}$$

and bound each of the three terms.

The next result is an upper bound on the first two terms.

Lemma 3.9.5. *Let $p \geq n \geq 4$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $\boldsymbol{\Sigma}$ is row-κ sparse. Consider any $\theta_{\vec{i},\vec{j}}$ that is a non-negative function of $\boldsymbol{u}_\ell$ for $\ell \in \vec{i} \cup \vec{j}$ defined for $\vec{i} \in C_q^<$ and $\vec{j} \in C_\delta^<$ with $1 \leq \delta \leq q \leq p-1$.*

(a) Suppose there exist positive constants a, z such that $\mathbb{E}\theta_{\vec{i},\vec{j}} \leq \mu_{n,q+1}(\boldsymbol{\Sigma}_{\vec{i}})az^q$ for any $\vec{j} \in J_{\vec{i}}$. Then

$$\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in J_{\vec{i}}}\mathbb{E}\theta_{\vec{i},\vec{j}} \leq ap(pz)^q\frac{q+1}{\delta!(q-\delta)!}\left(1+q^2\mu_{n,q+1}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)$$

(b) Suppose there exist positive constants a, z such that $\mathbb{E}\theta_{\vec{i},\vec{j}} \leq \mu_{n,q+\delta+2}(\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}})az^{q+\delta}$ for any $\vec{j} \in T_{\vec{i}}$. Then

$$\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in T_{\vec{i}}}\mathbb{E}\theta_{\vec{i},\vec{j}} \leq ap^2(pz)^{q+\delta}\frac{3}{\delta!(q-1)!}\left(1+\mu_{n,q+\delta+2}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right).$$

Proof:
(a) Since $|J_{\vec{i}}| = \binom{q+1}{1}\binom{q}{\delta}$

$$\begin{aligned}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in J_{\vec{i}}}\mathbb{E}\theta_{\vec{i},\vec{j}} \leq& az^q\binom{q+1}{1}\binom{q}{\delta}\sum_{\vec{i}\in C_q^<}\mu_{n,q+1}(\boldsymbol{\Sigma}_{\vec{i}})\\ \leq& ap(pz)^q\frac{q+1}{\delta!(q-\delta)!}\left(1+q^2\mu_{n,q+1}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right),\end{aligned}$$

where the last step follows from Lemma 3.9.3.

(b)

$$\begin{aligned}&\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in T_{\vec{i}}}\mathbb{E}\theta_{\vec{i},\vec{j}}\\ \leq& az^{q+\delta}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in T_{\vec{i}}}\mu_{n,q+\delta+2}(\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}})\end{aligned}$$

$$
\begin{aligned}
\leq & az^{q+\delta}\left(\sum_{\substack{\vec{i}\in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}}\sum_{\substack{\vec{j}\in T_{\vec{i}} \\ \mathbf{\Sigma}_{\vec{j}} \text{ diagonal}}} 1+\mu_{n,q+\delta+2}(\mathbf{\Sigma})\sum_{\substack{\vec{i}\in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}}\sum_{\substack{\vec{j}\in T_{\vec{i}} \\ \mathbf{\Sigma}_{\vec{j}} \text{ not diagonal}}} 1+\right. \\
& \left.\mu_{n,q+\delta+2}(\mathbf{\Sigma})\sum_{\substack{\vec{i}\in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}}\sum_{\vec{j}\in T_{\vec{i}}} 1\right) \\
\leq & az^{q+\delta}\left(\binom{p}{1}\binom{p-1}{q}\binom{p}{1}\binom{p-1}{\delta}+\mu_{n,q+\delta+2}(\mathbf{\Sigma})\binom{p}{1}\binom{p-1}{q}\sum_{\substack{\vec{j}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{j}} \text{ not diagonal}}} 1+\right. \\
& \left.\mu_{n,q+\delta+2}(\mathbf{\Sigma})\sum_{\substack{\vec{i}\in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}}\binom{p}{1}\binom{p-1}{\delta}\right) \\
\leq & ap^2(pz)^{q+\delta}\frac{3}{\delta!(q-1)!}\left(1+\mu_{n,q+\delta+2}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right),
\end{aligned}
$$

where the second inequality follows from that for $\vec{j}\in T_{\vec{i}}$, $\mathbf{\Sigma}_{\vec{i}\cup\vec{j}}$ is diagonal if and only if $\mathbf{\Sigma}_{\vec{i}}$ and $\mathbf{\Sigma}_{\vec{j}}$ are both diagonal; and the last step follows from Lemma 3.9.2.

□

To control $\mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in N_{\vec{i}}}\theta_{\vec{i},\vec{j}}$, we further partition $N_{\vec{i}}$ into 6 subsets as follows. For $\vec{i}\in C_q^<$ with $q\geq\delta$, define

$$
\begin{aligned}
\mathcal{K}_1\left(\vec{i}\right) &:= \left\{\vec{j}\in N_{\vec{i}} : j_0 = i_0\right\}, \\
\mathcal{K}_2\left(\vec{i}\right) &:= \left\{\vec{j}\in N_{\vec{i}} : j_0\neq i_0, j_0\in\bigcup_{\ell=1}^{q}\{i_\ell\}, i_0\in\bigcup_{\ell=1}^{\delta}\{j_\ell\}\right\}, \\
\mathcal{K}_3\left(\vec{i}\right) &:= \left\{\vec{j}\in N_{\vec{i}} : j_0\neq i_0, j_0\notin\bigcup_{\ell=1}^{q}\{i_\ell\}, i_0\in\bigcup_{\ell=1}^{\delta}\{j_\ell\}\right\}, \\
\mathcal{K}_4\left(\vec{i}\right) &:= \left\{\vec{j}\in N_{\vec{i}} : j_0\neq i_0, j_0\in\bigcup_{\ell=1}^{q}\{i_\ell\}, i_0\notin\bigcup_{\ell=1}^{\delta}\{j_\ell\}\right\}, \\
\mathcal{K}_5\left(\vec{i}\right) &:= \left\{\vec{j}\in N_{\vec{i}} : j_0\neq i_0, j_0\notin\bigcup_{\ell=1}^{q}\{i_\ell\}, i_0\notin\bigcup_{\ell=1}^{\delta}\{j_\ell\}, \left|\left(\bigcup_{\ell=1}^{q}\{i_\ell\}\right)\bigcap\left(\bigcup_{\ell=1}^{\delta}\{j_\ell\}\right)\right|\geq 1\right\},
\end{aligned}
$$

$$\mathcal{K}_6\left(\vec{i}\right) := \left\{\vec{j} \in N_{\vec{i}} : j_0 \neq i_0, j_0 \notin \bigcup_{\ell=1}^{q}\{i_\ell\}, i_0 \notin \bigcup_{\ell=1}^{\delta}\{j_\ell\}, \left(\bigcup_{\ell=1}^{q}\{i_\ell\}\right) \bigcap \left(\bigcup_{\ell=1}^{\delta}\{j_\ell\}\right) = \emptyset\right\}.$$

Then $N_{\vec{i}} = \cup_{w=1}^{6}\mathcal{K}_w\left(\vec{i}\right)$. Let $D_{\vec{i}}^m = \{\vec{j} \in N_{\vec{i}} : \left|(\cup_{\ell=1}^{q}\{i_\ell\}) \bigcap \left(\cup_{\ell=1}^{\delta}\{j_\ell\}\right)\right| = m\}$. We are now in a good position to present a lemma on $\mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in N_{\vec{i}}}\theta_{\vec{i},\vec{j}}$.

Lemma 3.9.6. *Let $p \geq n \geq 4$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $\boldsymbol{\Sigma}$ is row-κ sparse. Consider any $\theta_{\vec{i},\vec{j}}$ that is a non-negative function of $\boldsymbol{u}_\ell$ for $\ell \in \vec{i} \cup \vec{j}$ defined for $\vec{i} \in C_q^<$ and $\vec{j} \in C_\delta^<$ with $1 \leq \delta \leq q \leq p-1$. Suppose there exist positive constants a, b, z such that $\theta_{\vec{i},\vec{j}}$ satisfies:*

$$\begin{aligned}
\mathbb{E}\theta_{\vec{i},\vec{j}} &\leq \mu_{n,|\vec{i}\cup\vec{j}|}(\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}})az^{q+\delta-m}, \quad \forall\, \vec{j} \in \mathcal{K}_w\left(\vec{i}\right) \cap D_{\vec{i}}^m,\ \forall\, 0 \leq m \leq \delta-1,\ \forall w \in \{1,3,4\};\\
\mathbb{E}\theta_{\vec{i},\vec{j}} &\leq \mu_{n,|\vec{i}\cup\vec{j}|}(\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}})az^{q+\delta-m-1}, \quad \forall\, \vec{j} \in \mathcal{K}_2\left(\vec{i}\right) \cap D_{\vec{i}}^m,\ \forall\, 0 \leq m \leq \delta-2;\\
\mathbb{E}\theta_{\vec{i},\vec{j}} &\leq \mu_{n,|\vec{i}\cup\vec{j}|}(\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}})abz^{q+\delta-m}, \quad \forall\, \vec{j} \in \mathcal{K}_5\left(\vec{i}\right) \cap D_{\vec{i}}^m,\ \forall\, 1 \leq m \leq \delta;\\
\mathbb{E}\theta_{\vec{i},\vec{j}} &\leq \mu_{n,|\vec{i}\cup\vec{j}|}(\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}})az^{q+\delta}, \quad \forall\, \vec{j} \in \mathcal{K}_6\left(\vec{i}\right).
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in N_{\vec{i}}}\mathbb{E}\theta_{\vec{i},\vec{j}} \leq & ap(pz)^{q+1}\left(1+\mu_{n,q+\delta+1}\left(\boldsymbol{\Sigma}\right)\left(3q^2\right)\frac{\kappa-1}{p}\right)(1+pz)^{\delta-1}\,\delta\frac{4+b/z}{(\delta-1)!}+\\
& ap^2(pz)^{q+\delta}\frac{(\delta+1)(q+1)}{\delta!q!}\mu_{n,q+\delta+2}\left(\boldsymbol{\Sigma}\right)\frac{\kappa-1}{p}.
\end{aligned}$$

Proof: Since

$$N_{\vec{i}} = \bigcup_{w=1}^{6}\mathcal{K}_w\left(\vec{i}\right),$$

$$\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in N_{\vec{i}}}\mathbb{E}\theta_{\vec{i},\vec{j}} = \sum_{w=1}^{6}\sum_{\vec{i}\in C_q^<} I_w\left(\vec{i}\right), \tag{3.60}$$

with

$$I_w\left(\vec{i}\right) := \sum_{\vec{j}\in\mathcal{K}_w\left(\vec{i}\right)}\mathbb{E}\theta_{\vec{i},\vec{j}}.$$

Case 1: $p \geq q+\delta+2$
Obviously $\mathcal{K}_1\left(\vec{i}\right) = \bigcup_{m=0}^{\delta-1}\left(\mathcal{K}_1\left(\vec{i}\right) \cap D_{\vec{i}}^m\right)$. Then for any $\vec{i} \in C_q^<$ satisfying $\boldsymbol{\Sigma}_{\vec{i}}$ diagonal,

$$\left|\vec{j} \in \mathcal{K}_1\left(\vec{i}\right) \bigcap D_{\vec{i}}^m : \boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}} \text{ diagonal } \right|$$

$$
\begin{aligned}
&= \binom{q}{m} \frac{1}{(\delta-m)!} \sum_{j_1 \in [p] \setminus \mathcal{NZ}\left(\vec{i}\right)} \sum_{\substack{j_2 \in [p] \setminus \mathcal{NZ}\left(\vec{i}\right) \\ j_2 \notin \mathcal{NZ}(j_1)}} \sum_{\substack{j_3 \in [p] \setminus \mathcal{NZ}\left(\vec{i}\right) \\ j_3 \notin \cup_{\ell=1}^{2} \mathcal{NZ}(j_\ell)}} \cdots \sum_{\substack{j_{\delta-m} \in [p] \setminus \mathcal{NZ}\left(\vec{i}\right) \\ j_{\delta-m} \notin \cup_{\ell=1}^{\delta-m-1} \mathcal{NZ}(j_\ell)}} 1 \\
&\geq \binom{q}{m} \frac{1}{(\delta-m)!} \prod_{\ell=0}^{\delta-m-1} \left(p_{\vec{i}} - \ell\kappa\right), \qquad (3.61)
\end{aligned}
$$

where in the first inequality we without loss of generality assume the components of $\vec{j}$ distinct from $\vec{i}$ are $j_1, j_2, \cdots, j_{\delta-m}$. Then

$$
\begin{aligned}
&\left| \vec{j} \in \mathcal{K}_1\left(\vec{i}\right) \bigcap D_{\vec{i}}^m : \mathbf{\Sigma}_{\vec{i} \cup \vec{j}} \text{ not diagonal} \right| \\
\leq &\binom{q}{m} \binom{p-1-q}{\delta-m} - \binom{q}{m} \frac{1}{(\delta-m)!} \prod_{\ell=0}^{\delta-m-1} \left(p_{\vec{i}} - \ell\kappa\right) \\
= &\binom{q}{m} \frac{1}{(\delta-m)!} \left(\prod_{\ell=0}^{\delta-m-1} (p-1-q-\ell) - \prod_{\ell=0}^{\delta-m-1} \left(p_{\vec{i}} - \ell\right) + \prod_{\ell=0}^{\delta-m-1} \left(p_{\vec{i}} - \ell\right) - \prod_{\ell=0}^{\delta-m-1} \left(p_{\vec{i}} - \ell\kappa\right) \right) \\
\leq &\binom{q}{m} \frac{1}{(\delta-m)!} (\delta-m) p^{\delta-m-1} (q+\delta)(\kappa-1), \qquad (3.62)
\end{aligned}
$$

where the first inequality follows from (3.61), and the second inequality follows from Lemma 3.14.3 (a), (b) and (3.79).

Then

$$
\begin{aligned}
&\sum_{\vec{i} \in C_q^<} I_1\left(\vec{i}\right) \\
= &\sum_{m=0}^{\delta-1} \left(\sum_{\substack{\vec{i} \in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}} \sum_{\substack{\vec{j} \in \mathcal{K}_1\left(\vec{i}\right) \bigcap D_{\vec{i}}^m \\ \mathbf{\Sigma}_{\vec{i} \cup \vec{j}} \text{ diagonal}}} + \sum_{\substack{\vec{i} \in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}} \sum_{\substack{\vec{j} \in \mathcal{K}_1\left(\vec{i}\right) \bigcap D_{\vec{i}}^m \\ \mathbf{\Sigma}_{\vec{i} \cup \vec{j}} \text{ not diagonal}}} + \sum_{\substack{\vec{i} \in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} \sum_{\vec{j} \in \mathcal{K}_1\left(\vec{i}\right) \bigcap D_{\vec{i}}^m} \right) \mathbb{E}\theta_{\vec{i},\vec{j}} \\
\leq &\sum_{m=0}^{\delta-1} \left(\sum_{\substack{\vec{i} \in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}} \sum_{\substack{\vec{j} \in \mathcal{K}_1\left(\vec{i}\right) \bigcap D_{\vec{i}}^m \\ \mathbf{\Sigma}_{\vec{i} \cup \vec{j}} \text{ diagonal}}} + \right. \\
&\left. \mu_{n,q+\delta+1}\left(\mathbf{\Sigma}\right) \left(\sum_{\substack{\vec{i} \in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}} \sum_{\substack{\vec{j} \in \mathcal{K}_1\left(\vec{i}\right) \bigcap D_{\vec{i}}^m \\ \mathbf{\Sigma}_{\vec{i} \cup \vec{j}} \text{ not diagonal}}} + \sum_{\substack{\vec{i} \in C_q^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} \sum_{\vec{j} \in \mathcal{K}_1\left(\vec{i}\right) \bigcap D_{\vec{i}}^m} \right) \right) a z^{q+\delta-m}
\end{aligned}
$$

$$
\begin{aligned}
\leq & \sum_{m=0}^{\delta-1} \left(\binom{p}{1}\binom{p-1}{q}\binom{q}{m}\binom{p-1-q}{\delta-m} + \right. \\
& \mu_{n,q+\delta+1}(\mathbf{\Sigma}) \binom{p}{1}\binom{p-1}{q}\binom{q}{m} \frac{\delta-m}{(\delta-m)!} p^{\delta-m-1}(q+\delta)(\kappa-1) \\
& \left. + \mu_{n,q+\delta+1}(\mathbf{\Sigma}) \frac{(q+1)}{2((q-1)!)} p^q(\kappa-1)\binom{q}{m}\binom{p-1-q}{\delta-m} \right) a z^{q+\delta-m} \\
\leq & \sum_{m=0}^{\delta-1} \left(\frac{1}{m!(\delta-m)!} + \mu_{n,q+\delta+1}(\mathbf{\Sigma}) \frac{\kappa-1}{p} \left(\frac{(q+\delta)(\delta-m)}{m!(\delta-m)!} + \frac{q(q+1)}{m!(\delta-m)!2} \right) \right) a p (pz)^{q+\delta-m} \\
\leq & a p (pz)^{q+1} \left(1 + \mu_{n,q+\delta+1}(\mathbf{\Sigma})(3q^2)\frac{\kappa-1}{p} \right) \frac{1}{(\delta-1)!} \sum_{m=0}^{\delta-1} \frac{(\delta-1)!}{m!(\delta-1-m)!} (pz)^{\delta-1-m} \\
= & a p (pz)^{q+1} \left(1 + \mu_{n,q+\delta+1}(\mathbf{\Sigma})(3q^2)\frac{\kappa-1}{p} \right) \frac{1}{(\delta-1)!} (1+pz)^{\delta-1}, \qquad (3.63)
\end{aligned}
$$

where the first inequality follows from $\mu_{n,q+\delta-m+1}(\mathbf{\Sigma}) \leq \mu_{n,q+\delta+1}(\mathbf{\Sigma})$, and the second inequality follows from Lemma 3.9.2 and (3.62).

Obviously

$$
\begin{aligned}
\mathcal{K}_2\left(\vec{i}\right) &= \bigcup_{m=0}^{\delta-2} \left(\mathcal{K}_2\left(\vec{i}\right) \bigcap D_{\vec{i}}^m \right), \quad \mathcal{K}_3\left(\vec{i}\right) = \bigcup_{m=0}^{\delta-1} \left(\mathcal{K}_3\left(\vec{i}\right) \bigcap D_{\vec{i}}^m \right), \\
\mathcal{K}_4\left(\vec{i}\right) &= \bigcup_{m=0}^{\delta-1} \left(\mathcal{K}_4\left(\vec{i}\right) \bigcap D_{\vec{i}}^m \right), \quad \mathcal{K}_5\left(\vec{i}\right) = \bigcup_{m=1}^{\delta} \left(\mathcal{K}_5\left(\vec{i}\right) \bigcap D_{\vec{i}}^m \right).
\end{aligned}
$$

Then following a similar analysis to $\mathcal{K}_1(\vec{i})$, additionally with Lemma 3.9.4, one obtain

$$
\begin{aligned}
\sum_{\vec{i}\in C_q^<} I_2\left(\vec{i}\right) \leq & a p (pz)^{q+1} \left(1 + \mu_{n,q+\delta}(\mathbf{\Sigma})(3q^2)\frac{\kappa-1}{p} \right) \frac{1}{(\delta-2)!} (1+pz)^{\delta-2} 1(\delta \geq 2), \qquad (3.64) \\
\sum_{\vec{i}\in C_q^<} I_3\left(\vec{i}\right) \leq & a p (pz)^{q+1} \left(1 + \mu_{n,q+\delta+1}(\mathbf{\Sigma})(3q^2)\frac{\kappa-1}{p} \right) \frac{1}{(\delta-1)!} (1+pz)^{\delta-1}, \qquad (3.65) \\
\sum_{\vec{i}\in C_q^<} I_4\left(\vec{i}\right) \leq & a p (pz)^{q+1} \left(1 + \mu_{n,2\delta+1}(\mathbf{\Sigma})(3q^2)\frac{\kappa-1}{p} \right) \frac{1}{(\delta-1)!} (1+pz)^{\delta-1}, \qquad (3.66) \\
\sum_{\vec{i}\in C_q^<} I_5\left(\vec{i}\right) \leq & a \left(\frac{b}{z}\right) p (pz)^{q+1} \left(1 + \mu_{n,q+\delta+1}(\mathbf{\Sigma})(3\delta^2)\frac{\kappa-1}{p} \right) \frac{1}{(\delta-1)!} (1+pz)^{\delta-1}. \qquad (3.67)
\end{aligned}
$$

The detailed derivation of the above inequalities are omitted for clean presentation.

Observe

$$\mathcal{K}_6\left(\vec{i}\right)=\left\{\vec{j}\in C_q^<:\left(\bigcup_{\ell=0}^{q}\{i_\ell\}\right)\bigcap\left(\bigcup_{\ell=0}^{\delta}\{j_\ell\}\right)=\emptyset,\ \exists\ell\in[\delta]\cup\{0\}\text{ such that }j_\ell\in\mathcal{NZ}\left(\vec{i}\right)\right\}.$$

Then

$$\begin{aligned}\left|\mathcal{K}_6\left(\vec{i}\right)\right|=&\binom{p-1-q}{1}\binom{p-2-q}{\delta}-\binom{p_{\vec{i}}}{1}\binom{p_{\vec{i}}-1}{\delta}\\ \leq&\frac{1}{\delta!}(\delta+1)p^\delta(q+1)(\kappa-1),\end{aligned}\tag{3.68}$$

where the inequality follows from Lemma 3.14.3 (a) and (3.79). Thus

$$\begin{aligned}\sum_{\vec{i}\in C_q^<}I_6\left(\vec{i}\right)\leq&\binom{p}{1}\binom{p-1}{q}\frac{1}{\delta!}(\delta+1)p^\delta(q+1)(\kappa-1)\mu_{n,q+\delta+2}\left(\mathbf{\Sigma}\right)az^{q+\delta},\\ \leq&ap^2(pz)^{q+\delta}\frac{(\delta+1)(q+1)}{\delta!q!}\mu_{n,q+\delta+2}\left(\mathbf{\Sigma}\right)\frac{\kappa-1}{p},\end{aligned}\tag{3.69}$$

where the first inequality follows from (3.68) and Lemma 3.3.3 (a).

Case 2: $p<q+\delta+2$

We have impose the condition $p\geq 2\delta+2$ to derive (3.63), (3.64), (3.65), (3.66), (3.67) and (3.69). However, one can verify directly these inequalities also holds when $p<q+\delta+2$. We omit these tedious verifications here and take it for granted (3.63), (3.64), (3.65), (3.66), (3.67) and (3.69) holds for all $1\leq\delta\leq q\leq p-1$.

Thus combining (3.60), (3.63), (3.64), (3.65), (3.66), (3.67) and (3.69), yield

$$\begin{aligned}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in N_{\vec{i}}}\mathbb{E}\theta_{\vec{i},\vec{j}}\leq&ap(pz)^{q+1}\left(1+\mu_{n,q+\delta+1}\left(\mathbf{\Sigma}\right)(3q^2)\frac{\kappa-1}{p}\right)(1+pz)^{\delta-1}\delta\frac{4+b/z}{(\delta-1)!}+\\ &ap^2(pz)^{q+\delta}\frac{(\delta+1)(q+1)}{\delta!q!}\mu_{n,q+\delta+2}\left(\mathbf{\Sigma}\right)\frac{\kappa-1}{p}.\end{aligned}$$

□

Combining Lemma 3.9.5 and Lemma 3.9.6 immediately yields the following lemma.

Lemma 3.9.7. *Let $p\geq n\geq 4$ and $\boldsymbol{X}\sim\mathcal{VE}(\boldsymbol{\mu},\mathbf{\Sigma},\theta)$. Suppose $\mathbf{\Sigma}$ is row-κ sparse. Consider any $\theta_{\vec{i},\vec{j}}$ that is a non-negative function of $\boldsymbol{u}_\ell$ for $\ell\in\vec{i}\cup\vec{j}$ defined for $\vec{i}\in C_q^<$ and $\vec{j}\in C_\delta^<$ with $1\leq\delta\leq q\leq p-1$. Suppose there exist a,z,b such that all conditions in Lemma 3.9.5 and in Lemma 3.9.6 hold. Moreover suppose $b/z\leq c_{n,\delta,q}$ for some positive constant $c_{n,\delta,q}$ that depends*

only on n, q and δ. Then

$$\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in C_\delta^<}\theta_{\vec{i},\vec{j}} \le C_{n,q,\delta}\left(p^{1+\frac{1}{\delta}}z\right)^q\left(1+(p^{1+\frac{1}{\delta}}z)^\delta\right)(1+pz)^{\delta-1}\left(1+\mu_{n,q+\delta+2}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right)ap^{1-\frac{q}{\delta}}$$

3.9.3 Proof of Proposition 3.3.6

In what follows in this subsection, for the sake of clean presentation, we write $C^<$ for $C_\delta^<$, and write $\Phi_{\vec{i}}$ for $\Phi_{\vec{i}}^{(\boldsymbol{R})}$, for any $\vec{i}\in C^<$.

Proof of Proposition 3.3.6: Recall

$$N_{E_\delta}^{(\boldsymbol{R})} = \sum_{\vec{i}\in C_\delta^<}\prod_{j=1}^{\delta}\Phi_{i_0i_j}^{(\boldsymbol{R})} = \sum_{\vec{i}\in C_\delta^<}\Phi_{\vec{i}}.$$

To apply a Compound Poisson Approximation result, some additional notations are needed to be introduced. For $\vec{i}\in C^<$, let $S_{\vec{i}}$ be defined in (3.32), and let $T_{\vec{i}}$, $N_{\vec{i}}$ be defined respectively in (3.58), (3.59) with $q=\delta$. Here $T_{\vec{i}}$ is the set of indexes consisting of coordinates outside neighborhood of $\vec{i}$; $N_{\vec{i}}$ is the set of "correlated but not highly correlated" indexes, i.e. the set of indexes of which at least one component is in the neighborhood of $\vec{i}$, but excluding those sets of indexes of which the set of components are the same as that of $\vec{i}$. Denote

$$W_{\vec{i}} = \sum_{\vec{j}\in T_{\vec{i}}}\Phi_{\vec{j}}, \quad Z_{\vec{i}} = \sum_{\vec{j}\in N_{\vec{i}}}\Phi_{\vec{j}}, \tag{3.70}$$

and recall $U_{\vec{i}} = \sum_{\vec{j}\in S_{\vec{i}}}\Phi_{\vec{j}}^{(\boldsymbol{R})}$ is defined in (3.33). Then $W_{\vec{i}}$ is independent of $U_{\vec{i}}$ and $\Phi_{\vec{i}}$. Further denote

$$\begin{aligned}\lambda_0 &= \sum_{\vec{i}\in C^<}\mathbb{E}\left(\frac{\Phi_{\vec{i}}}{\Phi_{\vec{i}}+U_{\vec{i}}}1\left(\Phi_{\vec{i}}+U_{\vec{i}}\ge 1\right)\right),\\ \zeta_{0\ell} &= \frac{1}{\lambda_0\ell}\sum_{\vec{i}\in C^<}\mathbb{E}\left(\Phi_{\vec{i}}1\left(\Phi_{\vec{i}}+U_{\vec{i}}=\ell\right)\right), \quad \forall \ell\ge 1\end{aligned} \tag{3.71}$$

and a probability distribution $\boldsymbol{\zeta}_0$ on positive integers with $\boldsymbol{\zeta}_0(\ell)=\zeta_{0\ell}$. The mean of $\boldsymbol{\zeta}_0$ is $\mathbb{E}\boldsymbol{\zeta}_0 = \sum_{\ell\ge1}\ell\zeta_{0\ell}$. Moreover, let $b_1 = \sum_{\vec{i}\in C^<}\mathbb{E}\Phi_{\vec{i}}\mathbb{E}(\Phi_{\vec{i}}+U_{\vec{i}}+Z_{\vec{i}})$ and

$$b_2 = \sum_{\vec{i}\in C^<}\mathbb{E}\left(\Phi_{\vec{i}}Z_{\vec{i}}\right). \tag{3.72}$$

In this proof we write λ and $\boldsymbol{\zeta}$ for $\lambda_{p,n,\delta,\rho}$ and $\boldsymbol{\zeta}_{n,\delta,\rho}(\ell)$ respectively when there is no confusion.

By the compound Poisson Stein's approximation, i.e. (5.19) and (5.16) in [Bar01],

$$d_{\mathrm{TV}}\left(\mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right), CP(\lambda, \boldsymbol{\zeta})\right) \leq e^{\lambda_0}\left(b_1 + b_2 + \lambda_0 d_W(\boldsymbol{\zeta}_0', \boldsymbol{\zeta}')\mathbb{E}\boldsymbol{\zeta}_0 + |\lambda_0 \mathbb{E}\boldsymbol{\zeta_0} - \lambda\mathbb{E}\boldsymbol{\zeta}|\right), \tag{3.73}$$

where $\boldsymbol{\zeta}_0'(\ell) = \ell\zeta_{0\ell}/\mathbb{E}\boldsymbol{\zeta}_0$ and $\boldsymbol{\zeta}'(\ell) = \ell\boldsymbol{\zeta}(\ell)/\mathbb{E}\boldsymbol{\zeta}$ for $\ell \in \mathbb{Z}_+$, the set of all positive integers. In (3.73), the distance d_W is the Wasserstein L_1 metric on probability measures over the set of positive integers $\mathbb{Z}_+$:

$$d_W(P, Q) = \sup_{f \in \mathrm{Lip}_1}\left|\int f dP - \int f dQ\right|$$

where $\mathrm{Lip}_1 = \{f : |f(r) - f(s)| \leq |r - s|, r, s \in \mathbb{Z}_+\}$.

By Lemma 3.14.2,

$$\begin{aligned}
\lambda_0 d_W(\boldsymbol{\zeta}_0', \boldsymbol{\zeta}')\mathbb{E}\boldsymbol{\zeta}_0 \leq & \lambda_0 \mathbb{E}\boldsymbol{\zeta}_0 \frac{\delta}{2}\sum_{\ell=1}^{\delta+1}\ell\left|\frac{\lambda_0\zeta_{0\ell}}{\lambda_0\mathbb{E}\boldsymbol{\zeta}_0} - \frac{\lambda\boldsymbol{\zeta}(\ell)}{\lambda\mathbb{E}\boldsymbol{\zeta}}\right| \\
= & \frac{\delta}{2}\sum_{\ell=1}^{\delta+1}\ell\left|(\lambda_0\zeta_{0\ell} - \lambda\boldsymbol{\zeta}(\ell)) + (\lambda\mathbb{E}\boldsymbol{\zeta} - \lambda_0\mathbb{E}\boldsymbol{\zeta}_0)\frac{\lambda\boldsymbol{\zeta}(\ell)}{\lambda\mathbb{E}\boldsymbol{\zeta}}\right| \\
\leq & \frac{\delta}{2}\sum_{\ell=1}^{\delta+1}\ell\left|\lambda_0\zeta_{0\ell} - \lambda\boldsymbol{\zeta}(\ell)\right| + \frac{\delta}{2}\left|\lambda\mathbb{E}\boldsymbol{\zeta} - \lambda_0\mathbb{E}\boldsymbol{\zeta}_0\right|\sum_{\ell=1}^{\delta+1}\ell\frac{\lambda\boldsymbol{\zeta}(\ell)}{\lambda\mathbb{E}\boldsymbol{\zeta}} \\
\leq & \delta\sum_{\ell=1}^{\delta+1}\ell\left|\lambda_0\zeta_{0\ell} - \lambda\boldsymbol{\zeta}(\ell)\right|.
\end{aligned}$$

Plug the above inequalities into (3.73),

$$d_{\mathrm{TV}}\left(\mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right), CP(\lambda, \boldsymbol{\zeta})\right) \leq e^{\lambda_0}\left(b_1 + b_2 + (\delta+1)\sum_{\ell=1}^{\delta+1}\ell\left|\lambda_0\zeta_{0\ell} - \lambda\boldsymbol{\zeta}(\ell)\right|\right). \tag{3.74}$$

It remains to estimate the quantities in the right hand side of (3.74).

Part I. Upper bound for λ_0 and $\sum_{\ell=1}^{\delta+1}\ell\left|\lambda_0\zeta_{0\ell} - \lambda\boldsymbol{\zeta}(\ell)\right|$

For $\ell \in [\delta + 1]$,

$$\begin{aligned}
& \left|\lambda_0\zeta_{0\ell} - \lambda\boldsymbol{\zeta}\left(\ell\right)\right| \\
= & \left|\lambda_0\zeta_{0\ell} - \frac{1}{\ell}\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\alpha(\ell, r_\rho)\right|
\end{aligned}$$

$$
\begin{aligned}
\leq & \frac{1}{\ell} \sum_{\vec{i} \in C^{<}} \left| \mathbb{E}\left(\Phi_{\vec{i}} 1\left(\Phi_{\vec{i}} + U_{\vec{i}} = \ell\right)\right) - (2P_n(r_\rho))^{\delta} \alpha(\ell, r_\rho) \right| \\
= & \frac{1}{\ell} \sum_{\substack{\vec{i} \in C^{<} \\ \boldsymbol{\Sigma}_{\vec{i}} \text{ not diagonal}}} \left| \mathbb{E}\left(\Phi_{\vec{i}} 1\left(\Phi_{\vec{i}} + U_{\vec{i}} = \ell\right)\right) - (2P_n(r_\rho))^{\delta} \alpha(\ell, r_\rho) \right| \\
\leq & \frac{1}{\ell} (\mu_{n,\delta+1}(\boldsymbol{\Sigma}) + 1)(2P_n(r_\rho))^{\delta} \alpha(\ell, r_\rho) \sum_{\substack{\vec{i} \in C^{<} \\ \boldsymbol{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1 \\
\leq & \frac{\alpha(\ell, r_\rho)}{\ell} \mu_{n,\delta+1}(\boldsymbol{\Sigma}) \gamma^{\delta} \frac{(\delta+1)}{(\delta-1)!} \frac{\kappa - 1}{p},
\end{aligned} \tag{3.75}
$$

where the first inequality follows from the definition of $\zeta_{0\ell}$ in (3.71), the second inequality follows from Lemma 3.9.1 and Lemma 3.3.3 (a), and the last inequality follows from Lemma 3.9.2 and $\mu_{n,\delta+1}(\boldsymbol{\Sigma}) \geq 1$.

Then

$$
|\lambda_0 \mathbb{E}\boldsymbol{\zeta_0} - \lambda \mathbb{E}\boldsymbol{\zeta}| \leq \sum_{\ell=1}^{\delta+1} \ell \left|\lambda_0 \zeta_{0\ell} - \lambda \boldsymbol{\zeta}(\boldsymbol{\ell})\right| \leq \mu_{n,\delta+1}(\boldsymbol{\Sigma}) \gamma^{\delta} \frac{(\delta+1)}{(\delta-1)!} \frac{\kappa - 1}{p}, \tag{3.76}
$$

where the last inequality follows from (3.75). As an immediate consequences,

$$
\begin{aligned}
\lambda_0 &\leq \lambda_0 \mathbb{E}\boldsymbol{\zeta_0} \\
&\leq |\lambda_0 \mathbb{E}\boldsymbol{\zeta_0} - \lambda \mathbb{E}\boldsymbol{\zeta}| + \lambda \mathbb{E}\boldsymbol{\zeta} \\
&\leq \mu_{n,\delta+1}(\boldsymbol{\Sigma}) \gamma^{\delta} \frac{(\delta+1)}{(\delta-1)!} \frac{\kappa - 1}{p} + \binom{p}{1} \binom{p-1}{\delta} (2P_n(r_\rho))^{\delta} \\
&\leq \gamma^{\delta} \frac{(\delta+1)}{(\delta-1)!} \left(\mu_{n,\delta+1}(\boldsymbol{\Sigma}) \frac{\kappa - 1}{p} + 1 \right),
\end{aligned} \tag{3.77}
$$

where the third inequality follows from (3.76).

Part II. Upper bound for b_1

Since $N_{\vec{i}} \cup S_{\vec{i}} \cup \{\vec{i}\} = C^{<} \backslash T_{\vec{i}}$,

$$
b_1 = \sum_{\vec{i} \in C^{<}} \sum_{\vec{j} \in C^{<} \backslash T_{\vec{i}}} \mathbb{E}\Phi_{\vec{i}} \mathbb{E}\Phi_{\vec{j}}. \tag{3.78}
$$

Given $\vec{i} \in C^<$, by (3.56) with $q = \delta$,

$$p_{\vec{i}} := \left| [p] \backslash \mathcal{NZ}\left(\vec{i}\right) \right| \geq p - (\delta+1)\kappa. \tag{3.79}$$

Since $|T_{\vec{i}}| = p_{\vec{i}} \binom{p_{\vec{i}}-1}{\delta}$,

$$|C^< \backslash T_{\vec{i}}| = p\binom{p-1}{\delta} - p_{\vec{i}}\binom{p_{\vec{i}}-1}{\delta} \leq \frac{1}{\delta!}(\delta+1)\left(\prod_{\alpha=0}^{\delta-1}(p-\alpha)\right)(p-p_{\vec{i}}) \leq \frac{(\delta+1)^2}{\delta!}p^\delta\kappa, \tag{3.80}$$

where the first inequality follows from Lemma 3.14.3 (a).

One straightforward upper bound is

$$\begin{aligned} b_1 &= \sum_{\vec{i}\in C^<} \sum_{\vec{j}\in C^< \backslash T_{\vec{i}}} \mathbb{E}\left(\prod_{l=1}^{\delta} \Phi^{(\boldsymbol{R})}_{i_0 i_l}\right) \mathbb{E}\left(\prod_{l'=1}^{\delta} \Phi^{(\boldsymbol{R})}_{j_0 j_{l'}}\right) \\ &\leq p\binom{p-1}{\delta}\frac{(\delta+1)^2}{\delta!}p^\delta\kappa\left(\mu_{n,\delta+1}\left(\boldsymbol{\Sigma}\right)\right)^2\left(2P_n(r_\rho)\right)^{2\delta} \\ &\leq \left(\mu_{n,\delta+1}\left(\boldsymbol{\Sigma}\right)\right)^2\frac{(\delta+1)^2}{(\delta!)^2}\left(2p^{1+\frac{1}{\delta}}P_n(r_\rho)\right)^{2\delta}\frac{\kappa}{p}, \end{aligned}$$

where the first inequality follows from Lemma 3.9.1, Lemma 3.3.3 (a) and (3.80). The $\left(\mu_{n,\delta+1}\left(\boldsymbol{\Sigma}\right)\right)^2$ in the above upper bound is not very satisfactory, and can be improved by a more involved analysis.

Observe for given $\vec{i} \in C^<$,

$$\begin{aligned} &\left|\{\vec{j} \in C^< \backslash T_{\vec{i}} : \boldsymbol{\Sigma}_{\vec{j}} \text{ not diagonal}\}\right| \\ =&|C^< \backslash T_{\vec{i}}| - \left(\left|\{\vec{j} \in C^< : \boldsymbol{\Sigma}_{\vec{j}} \text{ diagonal}\}\right| - \left|\{\vec{j} \in T_{\vec{i}} : \boldsymbol{\Sigma}_{\vec{j}} \text{ diagonal}\}\right|\right) \\ \leq&\frac{1}{\delta!}(\delta+1)\left(\prod_{\alpha=0}^{\delta-1}(p-\alpha)\right)(p-p_{\vec{i}}) - \left(\left|\{\vec{j} \in C^< : \boldsymbol{\Sigma}_{\vec{j}} \text{ diagonal}\}\right| - \left|\{\vec{j} \in T_{\vec{i}} : \boldsymbol{\Sigma}_{\vec{j}} \text{ diagonal}\}\right|\right), \end{aligned} \tag{3.81}$$

where the inequality follows from (3.80). Then

$$\begin{aligned} &\left|\{\vec{j} \in C^< : \boldsymbol{\Sigma}_{\vec{j}} \text{ diagonal}\}\right| - \left|\{\vec{j} \in T_{\vec{i}} : \boldsymbol{\Sigma}_{\vec{j}} \text{ diagonal}\}\right| \\ =&\frac{1}{\delta!}\sum_{j_0=1}^{p} \sum_{j_1\in[p]\backslash\mathcal{NZ}(j_0)} \cdots \sum_{j_\delta\in[p]\backslash\bigcup_{l=0}^{\delta-1}\mathcal{NZ}(j_l)} 1 - \frac{1}{\delta!} \sum_{\substack{j_0\in[p] \\ j_0\notin\mathcal{NZ}(\vec{i})}} \sum_{\substack{j_1\in[p]\backslash\mathcal{NZ}(j_0) \\ j_1\notin\mathcal{NZ}(\vec{i})}} \cdots \sum_{\substack{j_\delta\in[p]\backslash\bigcup_{l=0}^{\delta-1}\mathcal{NZ}(j_l) \\ j_\delta\notin\mathcal{NZ}(\vec{i})}} 1 \end{aligned} \tag{3.82}$$

$$
\begin{aligned}
=&\frac{1}{\delta!}\sum_{m=0}^{\delta}\Bigg(\sum_{\substack{j_0\in[p]\\ j_0\notin\mathcal{NZ}(\vec{i})}}\ \sum_{\substack{j_1\in[p]\setminus\mathcal{NZ}(j_0)\\ j_1\notin\mathcal{NZ}(\vec{i})}}\cdots \\
&\sum_{\substack{j_{m-1}\in[p]\setminus\bigcup_{l=0}^{m-2}\mathcal{NZ}(j_l)\\ j_{m-1}\notin\mathcal{NZ}(\vec{i})}}\ \sum_{\substack{j_m\in[p]\setminus\bigcup_{l=0}^{m-1}\mathcal{NZ}(j_l)\\ j_m\in\mathcal{NZ}(\vec{i})}}\ \sum_{j_{m+1}\in[p]\setminus\bigcup_{l=0}^{m}\mathcal{NZ}(j_l)}\cdots\sum_{j_\delta\in[p]\setminus\bigcup_{l=0}^{\delta-1}\mathcal{NZ}(j_l)}1\Bigg)\\
=&\frac{1}{\delta!}\sum_{\substack{j_0\in[p]\\ j_0\in\mathcal{NZ}(\vec{i})}}\ \sum_{j_1\in[p]\setminus\mathcal{NZ}(j_0)}\cdots\sum_{j_\delta\in[p]\setminus\bigcup_{l=0}^{\delta-1}\mathcal{NZ}(j_l)}1+\frac{1}{\delta!}\sum_{m=1}^{\delta}(\\
&\sum_{j_m\in\mathcal{NZ}(\vec{i})}\ \sum_{\substack{j_0\in[p]\\ j_0\notin\mathcal{NZ}(\vec{i})\\ j_0\notin\mathcal{NZ}(j_m)}}\ \sum_{\substack{j_1\in[p]\setminus\mathcal{NZ}(j_0)\\ j_1\notin\mathcal{NZ}(\vec{i})\\ j_1\notin\mathcal{NZ}(j_m)}}\cdots\sum_{\substack{j_{m-1}\in[p]\setminus\bigcup_{l=0}^{m-2}\mathcal{NZ}(j_l)\\ j_{m-1}\notin\mathcal{NZ}(\vec{i})\\ j_{m-1}\notin\mathcal{NZ}(j_m)}}\ \sum_{j_{m+1}\in[p]\setminus\bigcup_{l=0}^{m}\mathcal{NZ}(j_l)}\cdots\sum_{j_\delta\in[p]\setminus\bigcup_{l=0}^{\delta-1}\mathcal{NZ}(j_l)}1\Bigg)\\
\geq&\frac{1}{\delta!}(p-p_{\vec{i}})\prod_{\beta=1}^{\delta}(p-\beta\kappa)+\frac{1}{\delta!}\sum_{m=1}^{\delta}(p-p_{\vec{i}})\left(\prod_{\alpha=1}^{m}(p_{\vec{i}}-\alpha\kappa)\right)\left(\prod_{\beta=m+1}^{\delta}(p-\beta\kappa)\right)\\
\geq&\frac{(\delta+1)}{\delta!}(p-p_{\vec{i}})\prod_{\alpha=1}^{\delta}(p_{\vec{i}}-\alpha\kappa),
\end{aligned}
\tag{3.83}
$$

where the second equality follows by writing (3.82) as a telescoping sum with the convention that the summation over j_{-1} for $m=0$ and the summation over $j_{\delta+1}$ for $m=\delta$ vanish, and the third equality follows from changing the order of the summation for $m\geq 1$. Plug (3.83) into (3.81),

$$
\begin{aligned}
&\left|\{\vec{j}\in C^{<}\setminus T_{\vec{i}}:\mathbf{\Sigma}_{\vec{j}}\text{ not diagonal}\}\right|\\
\leq&\frac{1}{\delta!}(\delta+1)(p-p_{\vec{i}})\left(\prod_{\alpha=0}^{\delta-1}(p-\alpha)-\prod_{\alpha=0}^{\delta-1}(p_{\vec{i}}-1-\alpha)+\prod_{\alpha=1}^{\delta}(p_{\vec{i}}-\alpha)-\prod_{\alpha=1}^{\delta}(p_{\vec{i}}-\alpha\kappa)\right)\\
\leq&\frac{1}{\delta!}(\delta+1)(p-p_{\vec{i}})\left(\delta p^{\delta-1}(p-p_{\vec{i}}+1)+\frac{\delta(\delta+1)}{2}p^{\delta-1}(\kappa-1)\right)\\
\leq&\frac{3\delta(\delta+1)^3}{\delta!}p^{\delta-1}\kappa^2,
\end{aligned}
\tag{3.84}
$$

where the second inequality follows from Lemma 3.14.3 (a) and Lemma 3.14.3 (b), and the last

inequality follows from (3.79).

Then for any $\vec{i} \in C^<$,

$$\begin{aligned}\sum_{\vec{j}\in C^<\setminus T_{\vec{i}}} \mu_{n,\delta+1}\left(\mathbf{\Sigma}_{\vec{j}}\right) \leq & \sum_{\substack{\vec{j}\in C^<\setminus T_{\vec{i}} \\ \mathbf{\Sigma}_{\vec{j}} \text{ diagonal}}} 1+\mu_{n,\delta+1}\left(\mathbf{\Sigma}\right) \sum_{\substack{\vec{j}\in C^<\setminus T_{\vec{i}} \\ \mathbf{\Sigma}_{\vec{j}} \text{ not diagonal}}} 1 \\ \leq & \frac{(\delta+1)^2}{\delta!}p^{\delta}\kappa+\mu_{n,\delta+1}\left(\mathbf{\Sigma}\right)\frac{3\delta(\delta+1)^3}{\delta!}p^{\delta-1}\kappa^2, \\ \leq & \frac{3\delta(\delta+1)^3}{\delta!}p^{\delta}\kappa\left(1+\mu_{n,\delta+1}\left(\mathbf{\Sigma}\right)\frac{\kappa}{p}\right), \end{aligned} \tag{3.85}$$

where the first inequality follows from Lemma 3.3.3 (a), and the second inequality follows from (3.80), (3.84).

Then following (3.78),

$$\begin{aligned} b_1 \leq & \sum_{\vec{i}\in C^<}\sum_{\vec{j}\in C^<\setminus T_{\vec{i}}} \mu_{n,\delta+1}\left(\mathbf{\Sigma}_{\vec{i}}\right)\mu_{n,\delta+1}\left(\mathbf{\Sigma}_{\vec{j}}\right)(2P_n(r_\rho))^{2\delta} \\ \leq & \frac{p^{\delta+1}}{\delta!}\left(1+\delta^2\mu_{n,\delta+1}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right)\frac{3\delta(\delta+1)^3}{\delta!}p^{\delta}\kappa\left(1+\mu_{n,\delta+1}\left(\mathbf{\Sigma}\right)\frac{\kappa}{p}\right)(2P_n(r_\rho))^{2\delta} \\ \leq & \left(3\frac{\delta^3(\delta+1)^3}{(\delta!)^2}(2p^{1+\frac{1}{\delta}}P_n(r_\rho))^{2\delta}\right)\frac{\kappa}{p}\left(1+\mu_{n,\delta+1}(\mathbf{\Sigma})\frac{\kappa}{p}\right)^2, \end{aligned} \tag{3.86}$$

where the first inequality follows from Lemma 3.9.1, the second inequality follows from Lemma 3.9.3 and (3.85).

Part III. Upper bound for b_2

Let $\mathcal{K}_w(\vec{i})$ and $D^m_{\vec{i}}$ be the same as in Subsection 3.9.2 with $q=\delta$. It is straightforward by Lemma 3.3.3 (c), Lemma 3.9.1 and Lemma 3.9.4 that the conditions in Lemma 3.9.6 with $q=\delta$ and $\theta_{\vec{i},\vec{j}} = \Phi_{\vec{i}}\Phi_{\vec{j}}$ are satisfied with $a=1$, $b = 2P_n(2r_\rho)1(\delta\geq 2)+2P_n(r_\rho)1(\delta=1)$ and $z=2P_n(r_\rho)$. Moreover, $b/z \leq 2^{n-2}1(\delta\geq 2)+1$ by Lemma 3.14.1 (d). Thus by Lemma 3.9.6 with $q=\delta$ and $\theta_{\vec{i},\vec{j}} = \Phi_{\vec{i}}\Phi_{\vec{j}}$,

$$\begin{aligned} b_2 \leq & p(2pP_n(r_\rho))^{\delta+1}\left(1+\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)(3\delta^2)\frac{\kappa-1}{p}\right)(1+2pP_n(r_\rho))^{\delta-1}\,\delta\frac{5+2^{n-1}1(\delta\geq 2)}{(\delta-1)!}+ \\ & \frac{(\delta+1)^2}{(\delta!)^2}p^2(2pP_n(r_\rho))^{2\delta}\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\frac{\kappa-1}{p}. \end{aligned} \tag{3.87}$$

$$\begin{aligned} \leq & \gamma^{\delta+1}p^{-\frac{1}{\delta}}\left(1+\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)(3\delta^2)\frac{\kappa-1}{p}\right)\left(1+\gamma p^{-\frac{1}{\delta}}\right)^{\delta-1}\delta\frac{5+2^{n-1}1(\delta\geq 2)}{(\delta-1)!}+ \\ & \frac{(\delta+1)^2}{(\delta!)^2}\gamma^{2\delta}\mu_{n,2\delta+2}\left(\mathbf{\Sigma}\right)\frac{\kappa-1}{p}, \end{aligned} \tag{3.88}$$

where the last step follows from $2p^{1+\frac{1}{\delta}}P_n(r_\rho) \le \gamma$.

By combining (3.74), (3.76), (3.77), (3.86), (3.88), together with the assumption $2p^{1+\frac{1}{\delta}}P_n(r_\rho) \le \gamma$,

$$\begin{aligned}
&d_{\text{TV}}\left(\mathscr{L}\left(N_{E_\delta}^{(\boldsymbol{R})}\right), CP(\lambda, \boldsymbol{\zeta})\right) \\
\le& C_{n,\delta,\gamma}\left(C'_{\delta,\gamma}\right)^{\mu_{n,\delta+1}(\boldsymbol{\Sigma})(\kappa-1)/p}\left(\mu_{n,2\delta+2}\left(\boldsymbol{\Sigma}\right)\kappa/p\left(1+\mu_{n,2\delta+2}\left(\boldsymbol{\Sigma}\right)\left(\kappa/p\right)^2\right)+p^{-\frac{1}{\delta}}\right),
\end{aligned}$$

where

$$C'_{\delta,\gamma} = \exp\left(\gamma^\delta \frac{\delta+1}{(\delta-1)!}\right), \tag{3.89}$$

and

$$C_{n,\delta,\gamma} = C\frac{\delta^6+\delta^2 2^{n-1}1(\delta\ge 2)}{\delta!}\gamma^{\delta+1}(1+\gamma)^\delta C'_{\delta,\gamma}. \tag{3.90}$$

□

3.10 Proofs in Subsection 3.3.4

3.10.1 Proof of Lemma 3.3.8

Proof of Lemma 3.3.8: $N_{\breve{V}_\delta}^{(\boldsymbol{R})} \le N_{V_\delta}^{(\boldsymbol{R})} \le N_{E_\delta}^{(\boldsymbol{R})}$ follows trivially from their definitions. It remains to show

$$N_{E_\delta}^{(\boldsymbol{R})} \le (\delta+1)N_{E_{\delta+1}}^{(\boldsymbol{R})} + N_{\breve{V}_\delta}^{(\boldsymbol{R})}. \tag{3.91}$$

To see this, consider $\delta \ge 2$ and any vertex i and denote its degree by m. If $m < \delta$, then it contributes zero to both sides of (3.91). If $m = \delta$, then it contributes 1 to both sides of (3.91). If $m > \delta$, it contributes $\binom{m}{\delta}$ to left hand side of (3.91), while contributes $(\delta+1)\binom{m}{\delta+1} = (m-\delta)\binom{m}{\delta}$. The above observation proves (3.91). The case $\delta = 1$ is similar and is omitted.

The above proof indeed applies to any graph and, in particular, the empirical partial correlation graph. So the second equation in the statement of the lemma holds. □

3.10.2 Proof of Proposition 3.3.9 (a)

By (3.37), it suffices to establish an upper bound on $\mathbb{E}N_{E_{\delta+1}}^{(\boldsymbol{R})}$.

Lemma 3.10.1. *Let $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $\boldsymbol{\Sigma}$ is row-κ sparse. Let $\ell \in [p-1]$. Then*

$$\mathbb{E}N_{E_\ell}^{(\boldsymbol{R})} \le \frac{1}{\ell!}\left(1+\ell^2\mu_{n,\ell+1}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)p\left(2pP_n(r_\rho)\right)^\ell.$$

Proof:

$$\begin{aligned}\mathbb{E}N_{E_\ell}^{(\boldsymbol{R})} &= \sum_{\vec{i}\in C_\ell^<} \mathbb{E}\prod_{j=1}^{\ell}\Phi_{i_0 i_j}^{(\boldsymbol{R})} \\ &\le \sum_{\vec{i}\in C_\ell^<} \mu_{n,\ell+1}\left(\boldsymbol{\Sigma}_{\vec{i}}\right)\left(2P_n(r_\rho)\right)^\ell \\ &\le \frac{1}{\ell!}\left(1+\ell^2\mu_{n,\ell+1}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)p\left(2pP_n(r_\rho)\right)^\ell,\end{aligned}$$

where the first inequality follows from Lemma 3.9.1, and the second inequality follows from Lemma 3.9.3. □

Proof of Proposition 3.3.9 (a): It follows from (3.37), Lemma 3.10.1 and Lemma 3.14.4. □

3.10.3 Proof of Proposition 3.3.9 (b)

Similar to (3.30) and (3.31), denote

$$\Phi_{\vec{i}}^{(\boldsymbol{R})} = \prod_{j=1}^{\delta}\Phi_{i_0 i_j}^{(\boldsymbol{R})} = 1\left(\bigcap_{j=1}^{\delta}\{\mathbf{dist}(\boldsymbol{u}_{i_0},\boldsymbol{u}_{i_j})\le r_\rho\}\right).$$

Then by definition

$$N_{E_\delta}^{(\boldsymbol{P})} = \sum_{\vec{i}\in C_\delta^<}\Phi_{\vec{i}}^{(\boldsymbol{P})}. \tag{3.92}$$

By (3.31) and (3.92),

$$\left|N_{E_\delta}^{(\boldsymbol{P})} - N_{E_\delta}^{(\boldsymbol{R})}\right| \le \sum_{\vec{i}\in C_\delta^<}|\Phi_{\vec{i}}^{(\boldsymbol{P})} - \Phi_{\vec{i}}^{(\boldsymbol{R})}|.$$

The next three lemmas establish upper bound on $|\Phi_{\vec{i}}^{(\boldsymbol{P})} - \Phi_{\vec{i}}^{(\boldsymbol{R})}|$.

We may suppose $\boldsymbol{\Sigma}$ is (τ,κ) sparse throughout this proof and the proof of Proposition 3.3.9 (c) since the conclusion is invariant to permutation of the variables by Remark 3.2.7. As a result, the $\boldsymbol{U}$-score may be partitioned into $\hat{\boldsymbol{U}}\in\mathbb{R}^{(n-1)\times\tau}$ consisting of the first τ columns and $\check{\boldsymbol{U}}\in\mathbb{R}^{(n-1)\times(p-\tau)}$ consisting the remaining $p-\tau$ columns.

Denote $[\tau]=\{1,2,\cdots,\tau\}$. Define a matrix $\check{\boldsymbol{B}}$ by

$$\check{\boldsymbol{B}} = \frac{n-1}{p-\tau}\check{\boldsymbol{U}}[\check{\boldsymbol{U}}]^T = \frac{n-1}{p-\tau}\sum_{i\in[p]\setminus[\tau]}^{p}\boldsymbol{u}_i\boldsymbol{u}_i^T. \tag{3.93}$$

Denote $\check{\boldsymbol{Q}} = \sqrt{n-1}\check{\boldsymbol{U}}$. Observe $\check{\boldsymbol{Q}}$ has exactly $p-\tau$ independent columns and each column

$\sqrt{n-1}\boldsymbol{u}_i \sim \text{unif}(\sqrt{n-1}S^{n-2})$. These observations immediately give us part (a) of the following.

Lemma 3.10.2. *Let $\{\boldsymbol{u}_\alpha\}_{\alpha=1}^p$ be columns of $\boldsymbol{U}$ defined in Section 3.3.1. Let $\check{\boldsymbol{B}}$ be defined as in equation* (3.93)*.*

(a) Suppose $\boldsymbol{\Sigma}$ is (τ, κ) sparse. $\check{\boldsymbol{B}} = \frac{1}{p-\tau}\boldsymbol{Q}\boldsymbol{Q}^T$, where $\boldsymbol{Q} \in R^{(n-1)\times(p-\tau)}$ has independent columns with each column distributed as unif$(\sqrt{n-1}S^{n-2})$.

(b) $|\lambda_{max}\left(\frac{p}{p-\tau}\boldsymbol{B}\right) - \lambda_{max}(\check{\boldsymbol{B}})| \leq \frac{n-1}{p-\tau}\tau$, and $\lambda_{min}\left(\frac{p}{p-\tau}\boldsymbol{B}\right) \geq \lambda_{min}(\check{\boldsymbol{B}})$.

Proof: (b) Recall $\boldsymbol{B} = \frac{n-1}{p}\sum_{i=1}^p \boldsymbol{u}_i\boldsymbol{u}_i^T$. Then,

$$\frac{p}{p-\tau}\boldsymbol{B} - \check{\boldsymbol{B}} = \frac{n-1}{p-\tau}\sum_{i\in[\tau]}\boldsymbol{u}_i\boldsymbol{u}_i^T.$$

By Lemma 3.14.5 (a), we have:

$$\begin{aligned}
|\lambda_{\max}\left(\frac{p}{p-\tau}\boldsymbol{B}\right) - \lambda_{\max}\left(\check{\boldsymbol{B}}\right)| &\leq \left\|\frac{n-1}{p-\tau}\sum_{i\in[\tau]}\boldsymbol{u}_i\boldsymbol{u}_i^T\right\|_2 \\
&\leq \frac{n-1}{p-\tau}\sum_{i\in[\tau]}\|\boldsymbol{u}_i\boldsymbol{u}_i^T\|_2 \\
&\leq \frac{n-1}{p-\tau}\tau,
\end{aligned}$$

where for the last inequality, we use the fact that $\boldsymbol{u}_i \in S^{n-2}$. Moreover, by Lemma 3.14.5 (c), we get $\lambda_{\min}\left(\frac{p}{p-\tau}\boldsymbol{B}\right) \geq \lambda_{\min}(\check{\boldsymbol{B}})$. □

Denote $h_0\left(\check{\boldsymbol{B}}\right) = \frac{\lambda_{\max}\left(\check{\boldsymbol{B}}\right)+\frac{n-1}{p-\tau}\tau}{\lambda_{\min}\left(\check{\boldsymbol{B}}\right)} = \frac{S_{\max}\left(\check{\boldsymbol{B}}\right)+\frac{n-1}{p-\tau}\tau}{S_{\min}\left(\check{\boldsymbol{B}}\right)}$ to be the perturbational condition number of $\check{\boldsymbol{B}}$, where $\lambda_{\max}\left(\check{\boldsymbol{B}}\right)$, $\lambda_{\min}\left(\check{\boldsymbol{B}}\right)$, $S_{\max}\left(\check{\boldsymbol{B}}\right)$, and $S_{\min}\left(\check{\boldsymbol{B}}\right)$ are respectively the largest eigenvalue, smallest eigenvalue, largest singular value and smallest singular value of $\check{\boldsymbol{B}}$.

Lemma 3.10.3. *Suppose $p \geq n$. Let $\{\boldsymbol{u}_\alpha\}_{\alpha=1}^p$ and $\{\boldsymbol{y}_\alpha\}_{\alpha=1}^p$ be defined as in Section 3.3.1. Consider distinct i, j satisfying $1 \leq i, j \leq p$. Then with probability* 1*,*

$$\frac{1}{h_0\left(\check{\boldsymbol{B}}\right)}\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 \leq \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 \leq h_0\left(\check{\boldsymbol{B}}\right)\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2,$$

and

$$\frac{1}{h_0\left(\check{\boldsymbol{B}}\right)}\|\boldsymbol{u}_i + \boldsymbol{u}_j\|_2 \leq \|\boldsymbol{y}_i + \boldsymbol{y}_j\|_2 \leq h_0\left(\check{\boldsymbol{B}}\right)\|\boldsymbol{u}_i + \boldsymbol{u}_j\|_2.$$

Proof: Recall $\boldsymbol{y}_\alpha = \bar{\boldsymbol{y}}_\alpha / \|\bar{\boldsymbol{y}}_\alpha\|_2$ and $\bar{\boldsymbol{y}}_\alpha = \boldsymbol{A}\boldsymbol{u}_\alpha$ $a.s.$, for $\alpha = i, j$. Apply the upper bound in Lemma 3.14.6,

$$\begin{aligned}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 &\leq \frac{\lambda_{\max}(\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A})}\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 \\ &= \frac{\lambda_{\max}\left(\frac{p}{p-\tau}\boldsymbol{B}\right)}{\lambda_{\min}\left(\frac{p}{p-\tau}\boldsymbol{B}\right)}\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 \quad a.s. \\ &\leq \frac{\lambda_{\max}\left(\check{\boldsymbol{B}}\right) + \frac{n-1}{p-\tau}\tau}{\lambda_{\min}\left(\check{\boldsymbol{B}}\right)}\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2,\end{aligned}$$

where the equality follows from the fact that $\boldsymbol{B} = \boldsymbol{A}^{-1}$ $a.s.$, and the last inequality follows from Lemma 3.10.2 (b). The lower bound of the first desired display follows similarly, by the lower bound in Lemma 3.14.6.

The second desired display follows analogously. □

For $\{i, j\} \in [p]$ with $i \neq j$, $q \in \{-1, +1\}$, define

$$\begin{aligned}&S_{ij}^{(q)}(r_\rho) = \{\|\boldsymbol{y}_i - q\boldsymbol{y}_j\|_2 \leq r_\rho\}, \quad F_{ij}^{(q)}(r_\rho) = \{\|\boldsymbol{u}_i - q\boldsymbol{u}_j\|_2 \leq r_\rho\}, \\ &G_{ij}^{(q)}(r_\rho) = \left\{\|\boldsymbol{u}_i - q\boldsymbol{u}_j\|_2 \leq \frac{1}{h_0\left(\check{\boldsymbol{B}}\right)} r_\rho\right\}, \quad H_{ij}^{(q)}(r_\rho) = \left\{\|\boldsymbol{u}_i - q\boldsymbol{u}_j\|_2 \leq h_0\left(\check{\boldsymbol{B}}\right) r_\rho\right\}.\end{aligned} \tag{3.94}$$

Define $F_{ij}(r_\rho) = F_{ij}^{(-1)}(r_\rho) \cup F_{ij}^{(+1)}(r_\rho)$. $G_{ij}(r_\rho)$, $H_{ij}(r_\rho)$, $S_{ij}(r_\rho)$ are defined similarly. Using these notations, then $\Phi_{ij}^{(\boldsymbol{P})}(\rho) = 1\left(S_{ij}(r_\rho)\right)$, and $\Phi_{ij}^{(\boldsymbol{R})}(\rho) = 1\left(F_{ij}(r_\rho)\right)$. For $\vec{i} \in C_\delta^<$, denote

$$H_{\vec{i}}(r_\rho) = \bigcap_{\ell=1}^{\delta} H_{i_0 i_\ell}(r_\rho), \quad H_{\vec{i},-m}(r_\rho) = \bigcap_{\substack{\ell=1 \\ \ell \neq m}}^{\delta} H_{i_0 i_\ell}(r_\rho).$$

When it's clear from the context, the dependence of r_ρ for the above quantities will be suppressed. By Lemma 3.10.3, with probability 1,

$$G_{ij}^{(q)} \subset S_{ij}^{(q)} \subset H_{ij}^{(q)}, \quad G_{ij}^{(q)} \subset F_{ij}^{(q)} \subset H_{ij}^{(q)}. \tag{3.95}$$

Lemma 3.10.4. *Suppose $p \geq n$. Consider $\delta \in [p-1]$. For any $\vec{i} \in C_\delta^<$, with probability* 1*,*

$$\left|\Phi_{\vec{i}}^{(\boldsymbol{P})} - \Phi_{\vec{i}}^{(\boldsymbol{R})}\right| \leq \xi_{\vec{i}},$$

where

$$\xi_{\vec{i}} := 1\left(\bigcup_{m=1}^{\delta}\left(\left(H_{i_0 i_m}\backslash G_{i_0 i_m}\right)\bigcap H_{\vec{i},-m}\right)\right).$$

Proof: Notice $\Phi_{\vec{i}}^{(\boldsymbol{R})} = 1\left(\bigcap_{m=1}^{\delta} F_{i_0 i_m}\right)$ and $\Phi_{\vec{i}}^{(\boldsymbol{P})} = 1\left(\bigcap_{m=1}^{\delta} S_{i_0 i_m}\right)$. Let $\triangle$ denote the symmetrization difference of two sets. Then

$$\left|\Phi_{\vec{i}}^{(\boldsymbol{P})} - \Phi_{\vec{i}}^{(\boldsymbol{R})}\right| = 1\left(\left(\bigcap_{m=1}^{\delta} F_{i_0 i_m}\right)\triangle\left(\bigcap_{m=1}^{\delta} S_{i_0 i_m}\right)\right) \leq \xi_{\vec{i}},$$

where the inequality follows from (3.95) and Lemma 3.14.7 (a). □

To control the expectation of the above term, we first bound the expectation on a high-probability set. Define the set $\mathcal{E}(t)$, with t being a parameter to be determined, by

$$\begin{aligned}\mathcal{E}(t) = &\left\{\left[1 - C_1\left(\sqrt{\frac{n-1}{p-\tau}} + \frac{t}{\sqrt{p-\tau}}\right)\right]^2 \leq \lambda_{\min}(\check{\boldsymbol{B}})\right\}\bigcap \\ &\bigcap\left\{\lambda_{\max}(\check{\boldsymbol{B}}) \leq \left[1 + C_1\left(\sqrt{\frac{n-1}{p-\tau}} + \frac{t}{\sqrt{p-\tau}}\right)\right]^2\right\},\end{aligned} \tag{3.96}$$

to be the set such that (3.177) in Lemma 3.14.8 holds, i.e. the constant C_1 in $\mathcal{E}(t)$ is the same constant as C in (3.177). By Lemma 3.10.2 (a) and Lemma 3.14.8,

$$\mathbb{P}(\mathcal{E}^c(t)) \leq 2\exp(-c_1 t^2). \tag{3.97}$$

Since $\tau \leq \frac{p}{2}$,

$$\frac{n-1}{p-\tau}\tau \leq 2(n-1)\frac{\tau}{p}, \tag{3.98}$$

and

$$C_1\left(\sqrt{\frac{n-1}{p-\tau}} + \frac{t}{\sqrt{p-\tau}}\right) \leq \sqrt{2}C_1\left(\sqrt{\frac{n-1}{p}} + \frac{t}{\sqrt{p}}\right). \tag{3.99}$$

Moreover, on $\mathcal{E}(t)$, and assuming

$$\sqrt{2}C_1\left(\sqrt{\frac{n-1}{p}} + \frac{t}{\sqrt{p}}\right) \leq \frac{1}{2}, \tag{3.100}$$

one has

$$h_0(\boldsymbol{B}) \leq \frac{\left(1+C_1\left(\sqrt{\frac{n-1}{p-\tau}}+\frac{t}{\sqrt{p-\tau}}\right)\right)^2+\frac{n-1}{p-\tau}\tau}{\left(1-C_1\left(\sqrt{\frac{n-1}{p-\tau}}+\frac{t}{\sqrt{p-\tau}}\right)\right)^2}$$
$$\leq 1+16\sqrt{2}C_1\left(\sqrt{\frac{n-1}{p}}+\frac{t}{\sqrt{p}}\right)+8(n-1)\frac{\tau}{p}:=\theta_1(t), \quad (3.101)$$

where the second inequality follows from (3.98), (3.99) and Lemma 3.14.3 (c).

For $\vec{i} \in C_\delta^<$, denote

$$F_{\vec{i}}(r_\rho)=\bigcap_{\ell=1}^{\delta} F_{i_0 i_\ell}(r_\rho), \quad F_{\vec{i},-m}(r_\rho)=\bigcap_{\substack{\ell=1\\ \ell\neq m}}^{\delta} F_{i_0 i_\ell}(r_\rho).$$

Lemma 3.10.5. *Let $p \geq n \geq 4$, $\delta \in [p-1]$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \leq \frac{p}{2}$. Let t be any positive number, and suppose* (3.100) *holds. Then for any $\vec{i} \in C_\delta^<$, with probability* 1*,*

$$\xi_{\vec{i}} 1\left(\mathcal{E}(t)\right) \leq \eta_{\vec{i}}(t),$$

where

$$\eta_{\vec{i}}(t) :=1\left(\bigcup_{m=1}^{\delta}\left(\left(F_{i_0 i_m}(\theta_1(t)r_\rho)\backslash F_{i_0 i_m}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\bigcap F_{\vec{i},-m}(\theta_1(t)r_\rho)\right)\right). \quad (3.102)$$

Moreover,

$$\mathbb{E}1\left(\left(F_{i_0 i_m}(\theta_1(t)r_\rho)\backslash F_{i_0 i_m}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\bigcap F_{\vec{i},-m}(\theta_1(t)r_\rho)\right)$$
$$\leq \mu_{n,\delta+1}(\boldsymbol{\Sigma}_{\vec{i}})2\left(P_n(r_\rho\theta_1(t))-P_n\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\left(2P_n\left(\theta_1(t)r_\rho\right)\right)^{\delta-1}, \quad (3.103)$$

and

$$\mathbb{E}\eta_{\vec{i}}(t) \leq \delta\mu_{n,\delta+1}(\boldsymbol{\Sigma}_{\vec{i}})2\left(P_n(r_\rho\theta_1(t))-P_n\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\left(2P_n\left(\theta_1(t)r_\rho\right)\right)^{\delta-1}$$
$$\leq \mu_{n,\delta+1}(\boldsymbol{\Sigma}_{\vec{i}})\delta n(\theta_1(t))^{n\delta}\left(\theta_1(t)-\frac{1}{\theta_1(t)}\right)\left(2P_n\left(r_\rho\right)\right)^{\delta}.$$

Proof: By (3.101), $H_{ij}(r_\rho) \cap \mathcal{E}(t) \subset F_{ij}(\theta_1(t)r_\rho)$ and $G_{ij}(r_\rho) \cap \mathcal{E}(t) \supset F_{ij}\left(\frac{r_\rho}{\theta_1(t)}\right)$. Then

$$\xi_{\vec{i}} 1\left(\mathcal{E}(t)\right) \leq \eta_{\vec{i}}(t). \quad (3.104)$$

$$
\begin{aligned}
&\mathbb{E}1\left(\left(F_{i_0i_m}(\theta_1(t)r_\rho)\backslash F_{i_0i_m}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\bigcap F_{\vec{i},-m}(\theta_1(t)r_\rho)\right)\\
\leq&\mu_{n,\delta+1}(\mathbf{\Sigma}_{\vec{i}})\mathbb{P}\left(\left(\bigcup_{q\in\{-1,+1\}}\left\{\frac{r_\rho}{\theta_1(t)}<\|\boldsymbol{u}'_{i_0}-q\boldsymbol{u}'_{i_m}\|_2\leq\theta_1(t)r_\rho\right\}\right)\bigcap\right.\\
&\left.\left(\bigcap_{\substack{\alpha=1\\\alpha\neq m}}^{\delta}\left(\bigcup_{q\in\{-1,+1\}}\{\|\boldsymbol{u}'_{i_0}-q\boldsymbol{u}'_{i_\alpha}\|_2\leq\theta_1(t)r_\rho\}\right)\right)\right),
\end{aligned}
\tag{3.105}
$$

where the last inequality follows from Lemma 3.3.3 (c) with

$$
\boldsymbol{u}'_{i_0},\boldsymbol{u}'_{i_1},\cdots,\boldsymbol{u}'_{i_\delta}\overset{i.i.d.}{\sim}\text{unif}(S^{n-2}).
$$

For any $\boldsymbol{w}\in S^{n-2}$, define $\Omega^{(q)}_{\boldsymbol{w}}:=\{\boldsymbol{v}\in S^{n-2}:\frac{1}{\theta_1(t)}r_\rho<\|\boldsymbol{v}-q\boldsymbol{w}\|_2\leq r_\rho\theta_1(t)\}$. Then

$$
\mathbb{P}\left(\boldsymbol{u}'_{im}\in\bigcup_{q\in\{-1,+1\}}\Omega^{(q)}_{\boldsymbol{w}}\right)=2\left(P_n(r_\rho\theta_1(t))-P_n\left(\frac{1}{\theta_1(t)}r_\rho\right)\right).
$$

By conditioning on $\boldsymbol{u}'_{i_0}$, the term in right hand side of (3.105) equals to

$$
\mu_{n,\delta+1}(\mathbf{\Sigma}_{\vec{i}})2\left(P_n(r_\rho\theta_1(t))-P_n\left(\frac{1}{\theta_1(t)}r_\rho\right)\right)\left(2P_n\left(\theta_1(t)r_\rho\right)\right)^{\delta-1},
$$

which then proves (3.103).

By union bound,

$$
\begin{aligned}
\mathbb{E}\eta_{\vec{i}}(t)\leq&\sum_{m=1}^{\delta}\mathbb{E}1\left(\left(F_{i_0i_m}(\theta_1(t)r_\rho)\backslash F_{i_0i_m}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\bigcap F_{\vec{i},-m}(\theta_1(t)r_\rho)\right)\\
\overset{(*)}{\leq}&\delta\mu_{n,\delta+1}(\mathbf{\Sigma}_{\vec{i}})2\left(P_n(r_\rho\theta_1(t))-P_n\left(\frac{1}{\theta_1(t)}r_\rho\right)\right)\left(2P_n\left(\theta_1(t)r_\rho\right)\right)^{\delta-1}\\
\overset{(**)}{\leq}&\delta\mu_{n,\delta+1}(\mathbf{\Sigma}_{\vec{i}})2(n-2)P_n(r_\rho)\left(\theta_1(t)\right)^{n-3}\left(\theta_1(t)-\frac{1}{\theta_1(t)}\right)\left(2P_n(r_\rho\theta_1(t))\right)^{\delta-1}\\
\overset{(***)}{\leq}&\delta\mu_{n,\delta+1}(\mathbf{\Sigma}_{\vec{i}})2(n-2)P_n(r_\rho)\left(\theta_1(t)\right)^{n-3}\left(\theta_1(t)-\frac{1}{\theta_1(t)}\right)\left(\left(\theta_1(t)\right)^{n-2}2P_n\left(r_\rho\right)\right)^{(\delta-1)},
\end{aligned}
$$

where $(*)$ follows from (3.103), $(**)$ follows from Lemma 3.14.1 (c), and $(***)$ follows from Lemma 3.14.1 (d).

□

Lemma 3.10.6. *Let $p \geq n \geq 4$, $\delta \in [p-1]$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Let t be any positive number, and suppose* (3.100) *holds. Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \leq \frac{p}{2}$. Then*

$$\left| N_{E_\delta}^{(\boldsymbol{P})} - N_{E_\delta}^{(\boldsymbol{R})} \right| 1\left(\mathcal{E}(t)\right) \leq \sum_{\vec{i} \in C_\delta^<} \eta_{\vec{i}}(t)$$

and

$$\mathbb{E} \sum_{\vec{i} \in C_\delta^<} \eta_{\vec{i}}(t) \leq \frac{Cn^2}{(\delta-1)!} \left(1 + \delta^2 \frac{\kappa - 1}{p} \mu_{n,\delta+1}(\boldsymbol{\Sigma})\right) \left(\theta_1(t)\right)^{n\delta} \left(\sqrt{\frac{1}{p}} + \frac{t}{\sqrt{p}} + \frac{\tau}{p}\right) p \left(2pP_n\left(r_\rho\right)\right)^\delta,$$

where C is an universal constant.

Proof:

$$\begin{aligned}
\left| N_{E_\delta}^{(\boldsymbol{P})} - N_{E_\delta}^{(\boldsymbol{R})} \right| 1\left(\mathcal{E}(t)\right) &\leq \sum_{\vec{i} \in C_\delta^<} \mathbb{E} \left| \Phi_{\vec{i}}^{(\boldsymbol{P})} - \Phi_{\vec{i}}^{(\boldsymbol{R})} \right| 1\left(\mathcal{E}(t)\right) \\
&\leq \sum_{\vec{i} \in C_\delta^<} \eta_{\vec{i}}(t),
\end{aligned}$$

where the last inequality follows from Lemma 3.10.4 and Lemma 3.10.5.

By Lemma 3.10.5,

$$\begin{aligned}
&\sum_{\vec{i} \in C_\delta^<} \mathbb{E}\eta_{\vec{i}} \\
\leq &\sum_{\vec{i} \in C_\delta^<} \mu_{n,\delta+1}(\boldsymbol{\Sigma}_{\vec{i}}) \delta n (\theta_1(t))^{n\delta} \left(\theta_1(t) - \frac{1}{\theta_1(t)}\right) \left(2P_n\left(r_\rho\right)\right)^\delta \\
\leq &\frac{p^{\delta+1}}{\delta!} \left(1 + \delta^2 \frac{\kappa - 1}{p} \mu_{n,\delta+1}(\boldsymbol{\Sigma})\right) \delta n (\theta_1(t))^{n\delta} \left(\theta_1(t) - \frac{1}{\theta_1(t)}\right) \left(2P_n\left(r_\rho\right)\right)^\delta \\
\leq &\frac{Cn}{(\delta-1)!} \left(1 + \delta^2 \frac{\kappa - 1}{p} \mu_{n,\delta+1}(\boldsymbol{\Sigma})\right) \left(\theta_1(t)\right)^{n\delta} \left(\sqrt{\frac{n}{p}} + \frac{t}{\sqrt{p}} + n\frac{\tau}{p}\right) \left(2p^{1+\frac{1}{\delta}} P_n\left(r_\rho\right)\right)^\delta, \quad (3.106)
\end{aligned}$$

where the third inequality follows from Lemma 3.9.3 and the last inequality follows from Lemma 3.14.3 (d) and (3.101). □

Lemma 3.10.7. *Let $p \geq n \geq 4$, $\delta \in [p-1]$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \leq \frac{p}{2}$. Suppose $2p^{1+\frac{1}{\delta}} P_n(r_\rho) \leq \gamma$ and*

$\left(\sqrt{\frac{n-1}{p}}+\sqrt{\frac{\delta\ln p}{p}}\right)\le c$ *hold for some positive and small enough universal constant* c*. Then*

$$\mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{P})}-N_{E_\delta}^{(\boldsymbol{R})}\right|\le C_{E_\delta}^{(\boldsymbol{P})}\left(1+\frac{\kappa-1}{p}\mu_{n,\delta+1}(\boldsymbol{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}}+\frac{\tau}{p}\right),$$

where $C_{E_\delta}^{(\boldsymbol{P})}$ *is defined in* (3.109)*.*

Proof:

$$\begin{aligned}\mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{P})}-N_{E_\delta}^{(\boldsymbol{R})}\right|&\le\mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{P})}-N_{E_\delta}^{(\boldsymbol{R})}\right|1\left(\mathcal{E}(t)\right)+\binom{p}{1}\binom{p-1}{\delta}\mathbb{P}(\mathcal{E}^c(t)),\\&\le\mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{P})}-N_{E_\delta}^{(\boldsymbol{R})}\right|1\left(\mathcal{E}(t)\right)+\frac{p^{\delta+1}}{\delta!}2\exp(-c_1t^2),\end{aligned}\tag{3.107}$$

where the first inequality follows from $0\le N_{E_\delta}^{(k)}\le\binom{p}{1}\binom{p-1}{\delta}$ for both $k=\boldsymbol{R}$ and $k=\boldsymbol{P}$, and the second inequality follows from (3.97).

Choose $t=c_\delta\sqrt{\ln p}$ with $c_\delta=\sqrt{\frac{5\delta}{2c_1}}\ge\sqrt{\left(\frac{3}{2}+\delta\right)/c_1}$ such that

$$2\exp(-c_1t^2)\le2\exp\left(-\left(\frac{3}{2}+\delta\right)\ln p\right)=\frac{2}{p^{\frac{3}{2}+\delta}}.$$

Moreover, for any $c<\frac{1}{2\max\left\{\sqrt{\frac{5}{2c_1}},1\right\}\sqrt{2}C_1}$,

$$\left(\sqrt{\frac{n-1}{p}}+\sqrt{\frac{\delta\ln p}{p}}\right)\le c$$

implies

$$\sqrt{2}C_1\left(\sqrt{\frac{n-1}{p}}+c_\delta\sqrt{\frac{\ln p}{p}}\right)\le\frac{1}{2},\tag{3.108}$$

which is (3.100) with $t=c_\delta\sqrt{\ln p}$. Then apply Lemma 3.10.6 with $t=c_\delta\sqrt{\ln p}$ to (3.107),

$$\begin{aligned}&\mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{P})}-N_{E_\delta}^{(\boldsymbol{R})}\right|\\\le&\frac{Cn^2}{(\delta-1)!}\left(1+\delta^2\frac{\kappa-1}{p}\mu_{n,\delta+1}(\boldsymbol{\Sigma})\right)\left(\theta_1\left(c_\delta\sqrt{\ln p}\right)\right)^{n\delta}\left(\sqrt{\frac{1}{p}}+\frac{\sqrt{\delta\ln p}}{\sqrt{p}}+\frac{\tau}{p}\right)\gamma^\delta+\frac{2}{\delta!\sqrt{p}}\\\le&\frac{Cn^2\sqrt{\delta}}{(\delta-1)!}\left(1+\delta^2\frac{\kappa-1}{p}\mu_{n,\delta+1}(\boldsymbol{\Sigma})\right)\left(\theta_1(c_\delta\sqrt{\ln p})\right)^{n\delta}\left(\frac{\sqrt{\ln p}}{\sqrt{p}}+\frac{\tau}{p}\right)\gamma^\delta+\frac{2}{\delta!\sqrt{p}}\\\le&C_{E_\delta}^{(\boldsymbol{P})}\left(1+\frac{\kappa-1}{p}\mu_{n,\delta+1}(\boldsymbol{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}}+\frac{\tau}{p}\right),\end{aligned}$$

where

$$\begin{aligned} C_{E_\delta}^{(\boldsymbol{P})} =& \frac{Cn^2\delta^{\frac{5}{2}}}{(\delta-1)!}\left(\theta_1\left(c_\delta\sqrt{\ln p}\right)\right)^{n\delta}\gamma^\delta + \frac{2}{\delta!\sqrt{\ln p}} \\ \leq& \frac{Cn^2\delta^{\frac{5}{2}}}{(\delta-1)!}\left(4n+5\right)^{n\delta}\gamma^\delta + \frac{2}{\delta!}, \end{aligned} \tag{3.109}$$

where the last step follows from $\theta_1\left(c_\delta\sqrt{\ln p}\right) \leq 9 + 4(n-1) = 4n+5$ by (3.108) and $\tau \leq p/2$. □

Proof of Proposition 3.3.9 (b): It follows directly from Lemma 3.10.7 and Lemma 3.14.4. □

3.10.4 Proof of Proposition 3.3.9 (c)

By Lemma 3.3.8,

$$N_{E_\delta}^{(\boldsymbol{P})} - (\delta+1)N_{E_{\delta+1}}^{(\boldsymbol{P})} - N_{E_\delta}^{(\boldsymbol{R})} \leq N_{\check{V}_\delta}^{(\boldsymbol{P})} - N_{\check{V}_\delta}^{(\boldsymbol{R})} \leq N_{E_\delta}^{(\boldsymbol{P})} - N_{E_\delta}^{(\boldsymbol{R})} + (\delta+1)N_{E_{\delta+1}}^{(\boldsymbol{R})},$$

which implies

$$\left|N_{\check{V}_\delta}^{(\boldsymbol{P})} - N_{\check{V}_\delta}^{(\boldsymbol{R})}\right| \leq \left|N_{E_\delta}^{(\boldsymbol{P})} - N_{E_\delta}^{(\boldsymbol{R})}\right| + (\delta+1)\left|N_{E_{\delta+1}}^{(\boldsymbol{P})} - N_{E_{\delta+1}}^{(\boldsymbol{R})}\right| + (\delta+1)N_{E_{\delta+1}}^{(\boldsymbol{R})}. \tag{3.110}$$

Lemma 3.10.8. *Let $p \geq n \geq 4$, $\delta \in [p-1]$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \leq \frac{p}{2}$. Suppose $2p^{1+\frac{1}{\delta}}P_n(r_\rho) \leq \gamma$ and $\left(\sqrt{\frac{n-1}{p}} + \sqrt{\frac{\ln p}{p}}\right) \leq c$ hold for some positive and small universal constant c. Then*

$$\mathbb{E}\left|N_{\check{V}_\delta}^{(\boldsymbol{P})} - N_{\check{V}_\delta}^{(\boldsymbol{R})}\right| \leq C_{\check{V}_\delta}^{(\boldsymbol{P})}\left(1 + \frac{\kappa-1}{p}\mu_{n,\delta+2}(\boldsymbol{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}} + \frac{\tau}{p} + p^{-\frac{1}{\delta}}\right)$$

where $C_{\check{V}_\delta}^{(\boldsymbol{P})}$ is defined in (3.114)*.*

Proof: Let $\mathcal{E}(t)$ be the same as in (3.96) with t to be determined. Consider $\delta \in [p-2]$.

$$\begin{aligned} &\mathbb{E}\left|N_{\check{V}_\delta}^{(\boldsymbol{P})} - N_{\check{V}_\delta}^{(\boldsymbol{R})}\right| \\ \leq& \mathbb{E}\left|N_{\check{V}_\delta}^{(\boldsymbol{P})} - N_{\check{V}_\delta}^{(\boldsymbol{R})}\right| 1\left(\mathcal{E}(t)\right) + p\mathbb{P}(\mathcal{E}^c(t)) \\ \leq& \mathbb{E}\left|N_{E_\delta}^{(\boldsymbol{P})} - N_{E_\delta}^{(\boldsymbol{R})}\right| 1\left(\mathcal{E}(t)\right) + (\delta+1)\mathbb{E}\left|N_{E_{\delta+1}}^{(\boldsymbol{P})} - N_{E_{\delta+1}}^{(\boldsymbol{R})}\right| 1\left(\mathcal{E}(t)\right) + \\ & (\delta+1)\mathbb{E}N_{E_{\delta+1}}^{(\boldsymbol{R})} + 2p\exp(-c_1t^2), \end{aligned} \tag{3.111}$$

where the first inequality follows from $0 \leq N_{\breve{V}_\delta}^{(k)} \leq p$ for $k = \boldsymbol{R}$ and $\boldsymbol{P}$, the second inequality follows from (3.110) and (3.97). If $\delta = p-1$, then

$$\begin{aligned}\mathbb{E}\left|N_{\breve{V}_{p-1}}^{(\boldsymbol{P})} - N_{\breve{V}_{p-1}}^{(\boldsymbol{R})}\right| \leq & \mathbb{E}\left|N_{\breve{V}_{p-1}}^{(\boldsymbol{P})} - N_{\breve{V}_{p-1}}^{(\boldsymbol{R})}\right| 1\left(\mathcal{E}(t)\right) + p\mathbb{P}(\mathcal{E}^c(t)) \\ \leq & \mathbb{E}\left|N_{E_{p-1}}^{(\boldsymbol{P})} - N_{E_{p-1}}^{(\boldsymbol{R})}\right| 1\left(\mathcal{E}(t)\right) + 2p\exp(-c_1 t^2),\end{aligned}$$

which shows (3.111) also holds for $\delta = p-1$ with the convention $N_{E_p}^{(\boldsymbol{P})} = N_{E_p}^{(\boldsymbol{R})} = 0$.

Choose $t = \sqrt{\frac{3}{c_1}\ln p} := c_2\sqrt{\ln p}$, such that $p\exp(-c_1 t^2) = \frac{1}{p^2}$. Moreover, for any $c < \frac{1}{2\max\{c_2,1\}\sqrt{2}C_1}$,

$$\left(\sqrt{\frac{n-1}{p}} + \sqrt{\frac{\ln p}{p}}\right) \leq c$$

implies

$$\sqrt{2}C_1\left(\sqrt{\frac{n-1}{p}} + c_2\sqrt{\frac{\ln p}{p}}\right) \leq \frac{1}{2}, \tag{3.112}$$

which is (3.100) with $t = c_2\sqrt{\ln p}$. With $t = c_2\sqrt{\ln p}$ Lemma 3.10.6 become:

$$\begin{aligned}&\mathbb{E}\left|N_{V_\delta}^{(\boldsymbol{P})} - N_{V_\delta}^{(\boldsymbol{R})}\right| 1\left(\mathcal{E}(t)\right) \\ \leq & \frac{Cn^2\delta^2}{(\delta-1)!}\left(1 + \frac{\kappa-1}{p}\mu_{n,\delta+1}(\boldsymbol{\Sigma})\right)\left(\theta_1(c_2\sqrt{\ln p})\right)^{n\delta}\left(\frac{\sqrt{\ln p}}{\sqrt{p}} + \frac{\tau}{p}\right)p\left(2pP_n\left(r_\rho\right)\right)^\delta, \end{aligned} \tag{3.113}$$

Then for $\delta \in [p-1]$ apply (3.113) with δ, $\delta+1$ and Lemma 3.10.1 to (3.111),

$$\begin{aligned}&\mathbb{E}\left|N_{\breve{V}_\delta}^{(\boldsymbol{P})} - N_{\breve{V}_\delta}^{(\boldsymbol{R})}\right| \\ \leq & \frac{Cn^2\delta^2}{(\delta-1)!}\left(1 + \frac{\kappa-1}{p}\mu_{n,\delta+2}(\boldsymbol{\Sigma})\right)\left(\theta_1\left(c_2\sqrt{\ln p}\right)\right)^{n\delta}\left(\frac{\sqrt{\ln p}}{\sqrt{p}} + \frac{\tau}{p}\right)p\left(2pP_n\left(r_\rho\right)\right)^\delta + \\ & \frac{Cn^2(\delta+1)^3}{\delta!}\left(1 + \frac{\kappa-1}{p}\mu_{n,\delta+2}(\boldsymbol{\Sigma})\right)\left(\theta_1\left(c_2\sqrt{\ln p}\right)\right)^{n(\delta+1)}\left(\frac{\sqrt{\ln p}}{\sqrt{p}} + \frac{\tau}{p}\right)p\left(2pP_n\left(r_\rho\right)\right)^{\delta+1} \\ & + (\delta+1)\frac{1}{(\delta+1)!}\left(1 + (\delta+1)^2\mu_{n,\delta+2}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)p\left(2pP_n(r_\rho)\right)^{\delta+1} + \frac{2}{p^2} \\ \leq & \frac{Cn^2(\delta+1)^3}{(\delta-1)!}\left(1 + \frac{\kappa-1}{p}\mu_{n,\delta+2}(\boldsymbol{\Sigma})\right)\left(\theta_1\left(c_2\sqrt{\ln p}\right)\right)^{n(\delta+1)}\left(\frac{\sqrt{\ln p}}{\sqrt{p}} + \frac{\tau}{p}\right)\gamma^\delta\left(1 + \gamma p^{-\frac{1}{\delta}}\right) \\ & + \frac{1}{\delta!}\left(1 + (\delta+1)^2\mu_{n,\delta+2}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)\gamma^{\delta+1}p^{-\frac{1}{\delta}} + \frac{2}{p^2} \\ \leq & C_{\breve{V}_\delta}^{(\boldsymbol{P})}\left(1 + \frac{\kappa-1}{p}\mu_{n,\delta+2}(\boldsymbol{\Sigma})\right)\left(\sqrt{\frac{\ln p}{p}} + \frac{\tau}{p} + p^{-\frac{1}{\delta}}\right),\end{aligned}$$

where the first inequality follows from additionally from $\mu_{n,\delta+1}(\mathbf{\Sigma}) \leq \mu_{n,\delta+2}(\mathbf{\Sigma})$, and in the last inequality

$$\begin{aligned} C_{\check{V}_\delta}^{(\boldsymbol{P})} =& \frac{Cn^2(\delta+1)^3}{(\delta-1)!}\left(\theta_1\left(c_2\sqrt{\ln p}\right)\right)^{n(\delta+1)}\gamma^\delta(1+\gamma)\left(1+\gamma p^{-\frac{1}{\delta}}\right)+\frac{2}{p^{2-\frac{1}{\delta}}} \\ \leq& \frac{Cn^2(\delta+1)^2}{(\delta-1)!}(4n+5)^{n(\delta+1)}\gamma^\delta(1+\gamma)\left(1+\frac{\delta+1}{\delta}\gamma p^{-\frac{1}{\delta}}\right)+2. \end{aligned} \tag{3.114}$$

where the last step follows from $\theta_1\left(c_2\sqrt{\ln p}\right) \leq 9+4(n-1) = 4n+5$ by (3.112) and $\tau \leq p/2$. □

Proof of Proposition 3.3.9 (c): The Lemma 3.10.8 and Lemma 3.14.4 complete the proof of Proposition 3.3.9 (c). □

3.11 Proofs in Subsection 3.3.5

3.11.1 Proof of Lemma 3.3.15

To utilize the notations we have defined in this chapter, we make the following adjustments on the notations throughout this subsection. In this proof it suffices to prove the conclusion for any $\delta+1$ i.i.d. random points from unif(S^{n-2}). Without loss of generality assume in this subsection that the first $\delta+1$ $\boldsymbol{U}$-scores $\{\boldsymbol{u}_i\}_{i=1}^{\delta+1}$ are independent. Another adjustment is to replace r by r_ρ. With these adjustments Lemma 3.3.15 is equivalent to prove: when $r_\rho < 2/\sqrt{5}$, $\delta \geq 1$, for any $\ell \in [\delta+1]$,

$$\begin{aligned} &\mathbb{P}\left(\textbf{NMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-1\right) = \ell | \deg(\boldsymbol{u}_{\delta+1}) = \delta\right) \\ =&\mathbb{P}\left(\textbf{PNMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-1\right) = \ell | \deg(\boldsymbol{u}_{\delta+1}) = \delta\right). \end{aligned} \tag{3.115}$$

Take $\vec{i} = (\delta+1, 1, \cdots, \delta)$. Recall the notation $\vec{i}, \Phi_{\vec{i}} = \Phi_{\vec{i}}^{(\boldsymbol{R})}, U_{\vec{i}}$ are defined in Subsection 3.3.3, where the dependence of $\boldsymbol{R}$ in $\Phi_{\vec{i}}^{(\boldsymbol{R})}$ is suppressed throughout this subsection for the sake of clean presentation. Then $\textbf{PNMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-1\right) = U_{\vec{i}} + \Phi_{\vec{i}}$. Moreover, the event $\{\deg(\boldsymbol{u}_{\delta+1}) = \delta\}$ in $\textbf{PGe}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-1\right)$ is the same as $\{\Phi_{\vec{i}} = 1\}$. Define $F_{ij}^{(q)} = \{\|\boldsymbol{u}_i - q\boldsymbol{u}_j\|_2 \leq r_\rho\}$. Then $1\left(\bigcap_{j=1}^{\delta} F_{j(\delta+1)}^{(+1)}\right)$ is the indicator function that the degree of vertex $\boldsymbol{u}_{\delta+1}$ in $\textbf{Ge}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-1\right)$ is δ. Hence Lemma 3.3.15 is equivalent to (3.115), which is

equivalent to: when $r_\rho < 2/\sqrt{5}$, $\delta \geq 1$, for any $\ell \in [\delta+1]$,

$$\mathbb{P}\left(U_{\vec{i}} + \Phi_{\vec{i}} = \ell | \Phi_{\vec{i}} = 1\right) = \mathbb{P}\left(\textbf{NMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1}, r_\rho; \delta+1, n-1\right) = \ell \middle| 1\left(\bigcap_{j=1}^{\delta} F_{j(\delta+1)}^{(+1)}\right) = 1\right). \tag{3.116}$$

Proof of (3.116)**:** For $\vec{q} = (q_1, q_2, \cdots, q_\delta) \in \{-1,+1\}^\delta$, denote $F_{\delta+1}^{(\vec{q})} = \bigcap_{j=1}^{\delta} F_{j(\delta+1)}^{(q_j)}$. Observe

$$\{\Phi_{\vec{i}} = 1\} = \bigcap_{j=1}^{\delta} \bigcup_{q_j \in \{+1,-1\}} F_{j(\delta+1)}^{(q_j)} == \bigcup_{\vec{q} \in \{-1,+1\}^\delta} F_{\delta+1}^{(\vec{q})}.$$

Since $r_\rho < 2/\sqrt{5} < \sqrt{2}$, $F_{j(\delta+1)}^{(-1)}$ and $F_{j(\delta+1)}^{(+1)}$ are disjoint for every $j \in [\delta]$, which implies $F_{\delta+1}^{(\vec{q})}$ for different $\vec{q} \in \{-1,+1\}^\delta$ are disjoint. Hence,

$$\mathbb{P}(\Phi_{\vec{i}} = 1, U_{\vec{i}} = \ell - 1) = \sum_{\vec{q} \in \{-1,+1\}^\delta} \mathbb{P}(F_{\delta+1}^{(\vec{q})}, U_{\vec{i}} = \ell - 1). \tag{3.117}$$

Next observe $1(F_{\delta+1}^{(\vec{q})})$ is a function of $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{\delta+1}$, and hence it has the same distribution as if replacing $\boldsymbol{u}_i$ by $-\boldsymbol{u}_i$ for any $i \in [\delta]$. Moreover, replacing $\boldsymbol{u}_i$ by $-\boldsymbol{u}_i$ for any $i \in [\delta]$ wouldn't change $U_{\vec{i}}$. As a result, (3.117) implies

$$\mathbb{P}(\Phi_{\vec{i}} = 1, U_{\vec{i}} = \ell - 1) = 2^\delta \mathbb{P}\left(F_{\delta+1}^{(\vec{q}_0)}, U_{\vec{i}} = \ell - 1\right), \tag{3.118}$$

where $\vec{q}_0 = (+1, +1, \cdots, +1)$ is the vector in $\mathbb{R}^\delta$ with all its components $+1$.

Consider $\omega \in F_{\delta+1}^{(\vec{q}_0)}$. Then $\Phi_{\vec{i}}(\omega) = 1$ or equivalently, $\Phi_{i(\delta+1)}^{(\boldsymbol{R})}(\omega) = 1$ for any $i \in [\delta]$. Then

$$U_{\vec{i}}(\omega) = \sum_{\vec{j} \in S_{\vec{i}}} \Phi_{\vec{j}}(\omega) = \sum_{i=1}^{\delta} \prod_{\substack{j=1 \\ j \neq i}}^{\delta+1} \Phi_{ij}^{(\boldsymbol{R})}(\omega) = \sum_{i=1}^{\delta} \prod_{\substack{j=1 \\ j \neq i}}^{\delta} \Phi_{ij}^{(\boldsymbol{R})}(\omega). \tag{3.119}$$

Since for any distinct $i, j \in [\delta]$, $\|\boldsymbol{u}_i(\omega) - \boldsymbol{u}_j(\omega)\|_2 \leq \|\boldsymbol{u}_i - \boldsymbol{u}_{\delta+1}(\omega)\|_2 + \|\boldsymbol{u}_i(\omega) - \boldsymbol{u}_{\delta+1}(\omega)\|_2 \leq 2r_\rho < 4/\sqrt{5}$, $\|\boldsymbol{u}_i(\omega) + \boldsymbol{u}_j(\omega)\|_2 = \sqrt{4 - \|\boldsymbol{u}_i(\omega) - \boldsymbol{u}_j(\omega)\|_2^2} > 2/\sqrt{5} > r_\rho$. Thus $\Phi_{ij}^{(\boldsymbol{R})}(\omega) = 1_{F_{ij}^{(+1)}}(\omega)$. That is, in the set $F_{\delta+1}^{(\vec{q}_0)}$, (3.119) become

$$U_{\vec{i}} = \sum_{i=1}^{\delta} \prod_{\substack{j=1 \\ j \neq i}}^{\delta} 1(F_{ij}^{(+1)}) = \textbf{NMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta}, r_\rho; \delta, n-1\right), \tag{3.120}$$

which implies

$$\begin{aligned}(\Phi_{\vec{i}}+U_{\vec{i}})\,1\left(F_{\delta+1}^{(\vec{q}_0)}\right)=&\left(1+\mathbf{NMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta},r_\rho;\delta+1,n-1\right)\right)1\left(F_{\delta+1}^{(\vec{q}_0)}\right)\\=&\mathbf{NMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1},r_\rho;\delta+1,n-1\right)1\left(F_{\delta+1}^{(\vec{q}_0)}\right).\end{aligned}\tag{3.121}$$

Thus

$$\begin{aligned}\mathbb{P}\left(U_{\vec{i}}+\Phi_{\vec{i}}=\ell|\Phi_{\vec{i}}=1\right)=&\frac{2^\delta\mathbb{P}\left(F_{\delta+1}^{(\vec{q}_0)},U_{\vec{i}}=\ell-1\right)}{\mathbb{P}\left(\Phi_{\vec{i}}=1\right)}\\=&\frac{2^\delta\mathbb{P}\left(F_{\delta+1}^{(\vec{q}_0)},U_{\vec{i}}=\ell-1\right)}{\left(2P_n(r_\rho)\right)^\delta}\\=&\frac{\mathbb{P}\left(F_{\delta+1}^{(\vec{q}_0)},\mathbf{NMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1},r_\rho;\delta+1,n-1\right)=\ell\right)}{\left(P_n(r_\rho)\right)^\delta}\\=&\mathbb{P}\left(\mathbf{NMD}\left(\{\boldsymbol{u}_i\}_{i=1}^{\delta+1},r_\rho;\delta+1,n-1\right)=\ell\,\middle|\,1\left(\bigcap_{j=1}^{\delta}F_{j(\delta+1)}^{(+1)}\right)=1\right),\end{aligned}$$

where the first equality follows from (3.118), the second equality follows from Lemma 3.9.1, and the third equality follows from (3.121).

□

3.11.2 Proofs of Lemma 3.3.16 and Lemma 3.3.17

Proof of Lemma 3.3.16: In the set $\{\deg(\boldsymbol{u}'_{\delta+1})=\delta\}$, it follows that

$$\mathbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta+1},r;\delta+1,n-1\right)=\mathbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)+1.$$

Thus

$$\begin{aligned}&\mathbb{P}\left(\deg(\boldsymbol{u}'_{\delta+1})=\delta,\mathbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta+1},r;\delta+1,n-1\right)=\ell\right)\\=&\mathbb{E}1\left(\{\deg(\boldsymbol{u}'_{\delta+1})=\delta\}\bigcap\{\mathbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\}\right)\\=&\mathbb{E}\left(\mathbb{E}\left(1\left(\{\deg(\boldsymbol{u}'_{\delta+1})=\delta\}\bigcap\{\mathbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\}\right)\left|\boldsymbol{u}'_{\delta+1}\right.\right)\right)\\=&\mathbb{E}\left(1\left(\{\deg(\boldsymbol{u}'_{\delta+1})=\delta\}\bigcap\{\mathbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\}\right)\left|\boldsymbol{u}'_{\delta+1}=\boldsymbol{v}_0\right.\right),\end{aligned}\tag{3.122}$$

where the last equality follows from that

$$\mathbb{E}\left(1\left(\{\deg(\boldsymbol{u}'_{\delta+1})=\delta\}\bigcap\{\mathbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\}\right)\left|\boldsymbol{u}'_{\delta+1}\right.\right)$$

as a random variable of $\boldsymbol{u}'_{\delta+1}$, due to rotation invariance property of the distribution unif(S^{n-2}), is degenerate to the constant

$$\mathbb{E}\left(1\left(\{\deg(\boldsymbol{u}'_{\delta+1})=\delta\}\bigcap\{\textbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\}\right)\left|\boldsymbol{u}'_{\delta+1}=\boldsymbol{v}_0\right.\right)$$

a.s. with $\boldsymbol{v}_0=(1,0,0,\cdots,0)\in S^{n-2}$.

Under $\boldsymbol{u}'_{\delta+1}=\boldsymbol{v}_0$, $1\left(\{\deg(\boldsymbol{u}'_{\delta+1})=\delta\}\right)=\prod_{i=1}^{\delta}1\left(\boldsymbol{u}'_i\in\text{SC}(r,\boldsymbol{v}_0)\right)$, where $\text{SC}(r,\boldsymbol{v}_0)$ is defined in (3.27). Use the following coordinate system for each $\boldsymbol{u}'_i=\left(u'_{ji}:1\le j\le n-1\right)^T$ in the region $\text{SC}(r,\boldsymbol{v}_0)$:

$$\begin{cases}u'_{1i}=1-\frac{r^2r_i^2}{2},\\ u'_{2i}=rr_i\sqrt{1-\frac{r^2r_i^2}{4}}\cos(\theta_{2i}),\\ \vdots & \text{for } 1\le i\le\delta,\\ u'_{ji}=rr_i\sqrt{1-\frac{r^2r_i^2}{4}}\cos(\theta_{ji})\prod_{m=2}^{j-1}\sin(\theta_{mi}),\\ \vdots\\ u'_{(n-2)i}=rr_i\sqrt{1-\frac{r^2r_i^2}{4}}\sin(\theta_{2i})\cdots\sin(\theta_{(n-3)i})\cos(\theta_{(n-2)i}),\\ u'_{(n-1)i}=rr_i\sqrt{1-\frac{r^2r_i^2}{4}}\sin(\theta_{2i})\cdots\sin(\theta_{(n-3)i})\sin(\theta_{(n-2)i}),\end{cases}$$

where for each $i\in[\delta]$:

$$r_i\in[0,1],\theta_{ji}\in[0,\pi]\text{ for }2\le j\le n-3\text{ and }\theta_{(n-2)i}\in[0,2\pi).\tag{3.123}$$

Then

$$\begin{aligned}&r^{-(n-2)\delta}\mathbb{E}\left(1\left(\{\deg(\boldsymbol{u}'_{\delta+1})=\delta\}\bigcap\{\textbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\}\right)\left|\boldsymbol{u}'_{\delta+1}=\boldsymbol{v}_0\right.\right)\\ =&r^{-(n-2)\delta}\mathbb{E}\prod_{i=1}^{\delta}1\left(\boldsymbol{u}'_i\in\text{SC}(r,\boldsymbol{v}_0)\right)1\left(\textbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\right)\\ \overset{(*)}{=}&\frac{r^{-(n-2)\delta}}{|S^{n-2}|^{\delta}}\int\cdots\int_{\Omega_0}1\left(\textbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\right)\\ &\times\prod_{i=1}^{\delta}\left(r^{n-2}r_i^{n-3}\left(1-\frac{r^2r_i^2}{4}\right)^{\frac{n-4}{2}}dr_i\prod_{j=2}^{n-2}\left(\sin^{n-2-j}(\theta_{ji})d\theta_{ji}\right)\right)\\ =&\frac{1}{|S^{n-2}|^{\delta}}\int\cdots\int_{\Omega_0}1\left(\textbf{NMD}\left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta},r;\delta,n-1\right)=\ell-1\right)\end{aligned}$$

$$\times \prod_{i=1}^{\delta} \left(r_i^{n-3} \left(1 - \frac{r^2 r_i^2}{4}\right)^{\frac{n-4}{2}} dr_i \prod_{j=2}^{n-2} \left(\sin^{n-2-j}(\theta_{ji}) d\theta_{ji}\right) \right), \tag{3.124}$$

where Ω_0 in equality $(*)$ is the region described in (3.123). Denote by $f(r)$ the integrand in (3.124). $f(r)$ is a function of r_i, θ_{ji} for $2 \le j \le n-2$ and $1 \le i \le \delta$, of which the dependences are suppressed.

Generally **NMD** $\left(\{\boldsymbol{v}_i\}_{i=1}^{\delta}, r; \delta, n-1\right)$ is a function of $\left(1\left(\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2 \le r\right) : 1 \le i < j \le \delta\right)$ and it does not depend on specific location of each vertices. In (3.124) **NMD** $\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta}, r; \delta, n-1\right)$ is further a function of $\left(1\left(\|\boldsymbol{u}_i' - \boldsymbol{u}_j'\|_2 < r\right) : 1 \le i < j \le \delta\right)$ since $\|\boldsymbol{u}_i' - \boldsymbol{u}_j'\|_2 = r$ contributes nothing to the integral due to the fact it happens with zero Lebesgue measure on Ω_0. Write

$$\begin{aligned}\textbf{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta}, r; \delta, n-1\right) =& g\left(1\left(\|\boldsymbol{u}_i' - \boldsymbol{u}_j'\|_2 < r\right) : 1 \le i < j \le \delta\right) \\ =& g\left(1\left(\frac{1}{r}\|\boldsymbol{u}_i' - \boldsymbol{u}_j'\|_2 < 1\right) : 1 \le i < j \le \delta\right).\end{aligned}$$

Intrinsically, $1\left(\|\boldsymbol{u}_i' - \boldsymbol{u}_j'\|_2 < r\right)$ is tells whether there is an edge between vertex i and j, and the function g is the function taking all edge information among δ vertices and output the number of vertices with maximal degree $\delta - 1$.

Then as $r \to 0^+$,

$$\begin{aligned}\lim_{r\to 0^+} f(r) =& \prod_{i=1}^{\delta} \left(r_i^{n-3} \prod_{j=2}^{n-2} \left(\sin^{n-2-j}(\theta_{ji})\right)\right) \\ & \times \lim_{r\to 0^+} 1\left(g\left(1\left(\frac{1}{r}\|\boldsymbol{u}_i' - \boldsymbol{u}_j'\|_2 < 1\right) : 1 \le i < j \le \delta\right) = \ell - 1\right).\end{aligned} \tag{3.125}$$

Observe

$$\begin{aligned}&\lim_{r\to 0^+} \left(\frac{1}{r}\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2\right)^2 \\ =& \left(r_i\cos(\theta_{2i}) - r_j\cos(\theta_{2j})\right)^2 + \sum_{q=3}^{n-2}\left(r_i\prod_{m=2}^{q-1}\sin(\theta_{mi})\cos(\theta_{qi}) - r_j\prod_{m=2}^{q-1}\sin(\theta_{mj})\cos(\theta_{qj})\right)^2 \\ &+ \left(r_i\prod_{m=2}^{n-2}\sin(\theta_{mi}) - r_j\prod_{m=2}^{n-2}\sin(\theta_{mj})\right)^2.\end{aligned} \tag{3.126}$$

On Ω_0, for $1 \le i \le \delta$, define

$$\begin{cases} \tilde{u}'_{1i} = r_i \cos(\theta_{2i}), \\ \tilde{u}'_{ji} = r_i \cos(\theta_{(j+1)i}) \prod_{m=2}^{j} \sin(\theta_{mi}), \text{ for } 2 \le j \le n-3 \\ \tilde{u}'_{(n-2)i} = r_i \prod_{m=2}^{n-2} \sin(\theta_{mi}), \end{cases} \tag{3.127}$$

and $\tilde{\boldsymbol{u}}'_i = (\tilde{u}'_{ji} : 1 \le j \le n-2) \in B^{n-2}$. Then by (3.126)

$$\lim_{r \to 0^+} \frac{1}{r} \|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 = \|\tilde{\boldsymbol{u}}'_i - \tilde{\boldsymbol{u}}'_j\|_2,$$

which, together with (3.125), imply

$$\begin{aligned} \lim_{r \to 0^+} f(r) =& \prod_{i=1}^{\delta} \left(r_i^{n-3} \prod_{j=2}^{n-2} \left(\sin^{n-2-j}(\theta_{ji}) \right) \right) 1 \left(g \left(1 \left(\|\tilde{\boldsymbol{u}}'_i - \tilde{\boldsymbol{u}}'_j\|_2 < 1 \right) : 1 \le i < j \le \delta \right) = \ell - 1 \right) \\ =& \prod_{i=1}^{\delta} \left(r_i^{n-3} \prod_{j=2}^{n-2} \left(\sin^{n-2-j}(\theta_{ji}) \right) \right) 1 \left(g \left(1 \left(\|\tilde{\boldsymbol{u}}'_i - \tilde{\boldsymbol{u}}'_j\|_2 \le 1 \right) : 1 \le i < j \le \delta \right) = \ell - 1 \right) \\ =& \prod_{i=1}^{\delta} \left(r_i^{n-3} \prod_{j=2}^{n-2} \left(\sin^{n-2-j}(\theta_{ji}) \right) \right) 1 \left(\mathbf{NMD} \left(\{\tilde{\boldsymbol{u}}'_i\}_{i=1}^{\delta}, 1; \delta, n-2 \right) = \ell - 1 \right). \end{aligned}$$

where the second equality holds $a.s.$ with respect to the Lebesgue measure on Ω_0,

Moreover, $|f(r)| \le 1$, which is integrable over the bounded set Ω_0. Apply Dominated Convergence Theorem to (3.124),

$$\begin{aligned} & \lim_{r \to 0^+} r^{-(n-2)\delta} \mathbb{E} \left(1 \left(\{\deg(\boldsymbol{u}'_{\delta+1}) = \delta\} \bigcap \{\mathbf{NMD} \left(\{\boldsymbol{u}'_i\}_{i=1}^{\delta}, r; \delta, n-1 \right) = \ell - 1\} \right) \left| \boldsymbol{u}'_{\delta+1} = \boldsymbol{v}_0 \right. \right) \\ =& \frac{1}{|S^{n-2}|^{\delta}} \int \cdots \int_{\Omega_0} 1 \left(\mathbf{NMD} \left(\{\tilde{\boldsymbol{u}}'_i\}_{i=1}^{\delta}, 1; \delta, n-2 \right) = \ell - 1 \right) \times \\ & \prod_{i=1}^{\delta} \left(r_i^{n-3} dr_i \prod_{j=2}^{n-2} \left(\sin^{n-2-j}(\theta_{ji}) d\theta_{ji} \right) \right) \\ =& \frac{|B^{n-2}|^{\delta}}{|S^{n-2}|^{\delta}} \mathbb{P} \left(\mathbf{NMD} \left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2 \right) = \ell - 1 \right), \end{aligned} \tag{3.128}$$

where the parametrization (3.127) and the region Ω_0 coincide with the spherical coordinates for B^{n-2}.

Thus

$$\begin{aligned}
&\lim_{r\to 0^+}\mathbb{P}\left(\mathbf{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1},r;\delta+1,n-1\right)=\ell\,\middle|\deg(\boldsymbol{u}_{\delta+1}')=\delta\right)\\
=&\lim_{r\to 0^+} r^{-(n-2)\delta}\mathbb{P}\left(\deg(\boldsymbol{u}_{\delta+1}')=\delta,\mathbf{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1},r;\delta+1,n-1\right)=\ell\right)\frac{r^{(n-2)\delta}}{\left(P_n(r)\right)^{\delta}}\\
=&\frac{|B^{n-2}|^{\delta}}{|S^{n-2}|^{\delta}}\mathbb{P}\left(\mathbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta},1;\delta,n-2\right)=\ell-1\right)\frac{1}{(c_n)^{\delta}}\\
=&\frac{1}{(c_n)^{\delta}}\frac{|B^{n-2}|^{\delta}}{|S^{n-2}|^{\delta}}\mathbb{P}\left(\mathbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta},1;\delta,n-2\right)=\ell-1\right)\\
=&\mathbb{P}\left(\mathbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta},1;\delta,n-2\right)=\ell-1\right),
\end{aligned}$$

where the second equality follows from (3.122), (3.128) and Lemma 3.14.1 (b). □

Proof of Lemma 3.3.17: By (3.44), Lemma 3.3.16 and (3.7),

$$\lim_{\rho\to 1^-}\alpha(\ell,r_\rho)=\lim_{r_\rho\to 0^+}\alpha(\ell,r_\rho)=\alpha_\ell,\quad \forall \ell\in[\delta+1], \tag{3.129}$$

and thus

$$\lim_{\rho\to 1^-}\boldsymbol{\zeta}_{n,\delta,\rho}(\ell)=\boldsymbol{\zeta}_{n,\delta}(\ell),\quad \forall \ell\in[\delta+1]. \tag{3.130}$$

By Lemma 3.14.1 (b)

$$\lim_{p\to\infty}2p^{1+\frac{1}{\delta}}P_n(r_\rho)=\lim_{p\to\infty}2c_np^{1+\frac{1}{\delta}}r_\rho^{n-2}=\lim_{p\to\infty}2^{\frac{n}{2}}c_np^{1+\frac{1}{\delta}}(1-\rho)^{\frac{n-2}{2}}=e_{n,\delta}.$$

Then the preceding display and (3.129) yield

$$\lim_{p\to\infty}\lambda_{p,n,\delta,\rho}=\lim_{p\to\infty}\frac{1}{\delta!}p^{\delta+1}(2P_n(r_\rho))^{\delta}\sum_{\ell=1}^{\delta+1}\frac{\alpha(\ell,r_\rho)}{\ell}=\lim_{p\to\infty}\frac{1}{\delta!}\left(e_{n,\delta}\right)^{\delta}\sum_{\ell=1}^{\delta+1}\frac{\alpha_\ell}{\ell}=\lambda_{n,\delta}(e_{n,\delta}). \tag{3.131}$$

(3.130) and (3.131) immediately yield the conclusion. □

3.12 Proofs in Section 3.4

3.12.1 Proofs of Lemma 3.4.1 and Proposition 3.4.2

Proof of Lemma 3.4.1:

$$\mathbb{E}N_{E_\delta}^{(\boldsymbol{R})}-\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^{\delta}=\sum_{\substack{\vec{i}\in C_\delta^<\\ \boldsymbol{\Sigma}_{\vec{i}}\text{ not diagonal}}}\left(\mathbb{E}\prod_{j=1}^{\delta}\Phi_{i_0i_j}^{(\boldsymbol{R})}-(2P_n(r_\rho))^{\delta}\right) \tag{3.132}$$

$$\leq (\mu_{n,\delta+1}(\mathbf{\Sigma}) - 1)(2P_n(r_\rho))^\delta \sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1,$$

where the first inequality follows from Lemma 3.3.3 (c) and Lemma 3.9.1. By (3.132),

$$\mathbb{E}N_{E_\delta}^{(\boldsymbol{R})} - \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta \geq -(2P_n(r_\rho))^\delta \sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1.$$

Combining the preceding two displays,

$$\begin{aligned}\left|\mathbb{E}N_{E_\delta}^{(\boldsymbol{R})} - \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\right| \leq & \max\{1, \mu_{n,\delta+1}(\mathbf{\Sigma}) - 1)\}(2P_n(r_\rho))^\delta \sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1 \\ \leq & \mu_{n,\delta+1}(\mathbf{\Sigma})(2P_n(r_\rho))^\delta \frac{(\delta+1)}{2((\delta-1)!)} p^\delta(\kappa-1) \\ \leq & \frac{(\delta+1)}{2((\delta-1)!)}\gamma^\delta \mu_{n,\delta+1}(\mathbf{\Sigma})\frac{\kappa-1}{p},\end{aligned}$$

where the second inequality follows from Lemma 3.9.2. □

Proof of Proposition 3.4.2: Recall for $\vec{i} \in C_\delta^<$, $\Phi_{\vec{i}}^{(\boldsymbol{R})}$ is defined in (3.30), $U_{\vec{i}}$ is defined in (3.33), and $Z_{\vec{i}}$ and $W_{\vec{i}}$ are defined in (3.70). Then

$$N_{E_\delta}^{(\boldsymbol{R})} = \sum_{\vec{i}\in C_\delta^<} \Phi_{\vec{i}}^{(\boldsymbol{R})} = \Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} + Z_{\vec{i}} + W_{\vec{i}}.$$

Then

$$\left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2 = \sum_{\vec{i}\in C_\delta^<} \Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} + Z_{\vec{i}} + W_{\vec{i}}\right). \tag{3.133}$$

Step 1:
Since $\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}$ takes value in $[\delta+1]\bigcup\{0\}$,

$$\begin{aligned}\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) = & \sum_{\ell=1}^{\delta+1} \ell\, \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} 1\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} = \ell\right) \\ = & \sum_{\ell=1}^{\delta+1} \ell\, \mathbb{P}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} = 1\right)\mathbb{P}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} = \ell | \Phi_{\vec{i}}^{(\boldsymbol{R})} = 1\right).\end{aligned}$$

For $\vec{i} \in C_\delta^<$ such that $\mathbf{\Sigma}_{\vec{i}}$ diagonal, $\mathbb{P}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} = 1\right) = (2P_n(r_\rho))^\delta$ by Lemma 3.9.1 and

$$\mathbb{P}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}} = \ell \middle| \Phi_{\vec{i}}^{(\boldsymbol{R})} = 1\right) = \alpha(\ell, r_\rho)$$

by (3.44). Thus in this case,

$$\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) = (2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho).$$

Moreover, when $\vec{i} \in C_\delta^<$ such that $\mathbf{\Sigma}_{\vec{i}}$ is not diagonal, by Lemma 3.3.3 (c)

$$\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) \leq \mu_{n,\delta+1}(\mathbf{\Sigma})(2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho).$$

Then by the preceding two displays,

$$\begin{aligned}
&\sum_{\vec{i} \in C_\delta^<} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) - \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho) \\
&= \sum_{\substack{\vec{i} \in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} \left(\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) - (2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho)\right) \qquad (3.134) \\
&\leq \left(\mu_{n,\delta+1}(\mathbf{\Sigma}) - 1\right)(2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho) \sum_{\substack{\vec{i} \in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1.
\end{aligned}$$

By (3.134),

$$\sum_{\vec{i} \in C_\delta^<} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) - \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta \geq -(2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho) \sum_{\substack{\vec{i} \in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1.$$

By combining the preceding two displays,

$$\left|\sum_{\vec{i} \in C_\delta^<} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) - \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho)\right|$$

$$
\begin{aligned}
&\leq \mu_{n,\delta+1}(\mathbf{\Sigma})(2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho) \sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} 1 \\
&\leq \mu_{n,\delta+1}(\mathbf{\Sigma})(2P_n(r_\rho))^\delta \sum_{\ell=1}^{\delta+1} \ell\alpha(\ell, r_\rho) \frac{(\delta+1)}{2((\delta-1)!)} p^\delta(\kappa-1) \\
&\leq \mu_{n,\delta+1}(\mathbf{\Sigma})(2p^{1+1/\delta}P_n(r_\rho))^\delta \frac{(\delta+1)^2}{2((\delta-1)!)} \frac{\kappa-1}{p} \qquad (3.135)
\end{aligned}
$$

where the second inequality follows from Lemma 3.9.2.

Step 2:

$$
\begin{aligned}
&\sum_{\vec{i}\in C_\delta^<} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} W_{\vec{i}} - \left(\binom{p}{1} \binom{p-1}{\delta} (2P_n(r_\rho))^\delta \right)^2 \\
&= \sum_{\vec{i}\in C_\delta^<} \sum_{\vec{j}\in T_{\vec{i}}} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} \Phi_{\vec{j}}^{(\boldsymbol{R})} - \sum_{\vec{i}\in C_\delta^<} \sum_{\vec{j}\in C_\delta^<} (2P_n(r_\rho))^{2\delta} \\
&= \sum_{\vec{i}\in C_\delta^<} \sum_{\vec{j}\in T_{\vec{i}}} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} \mathbb{E}\Phi_{\vec{j}}^{(\boldsymbol{R})} - \sum_{\vec{i}\in C_\delta^<} \sum_{\vec{j}\in T_{\vec{i}}} (2P_n(r_\rho))^{2\delta} - \sum_{\vec{i}\in C_\delta^<} \sum_{\vec{j}\in C_\delta^<\setminus T_{\vec{i}}} (2P_n(r_\rho))^{2\delta} \\
&= \sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} \sum_{\vec{j}\in T_{\vec{i}}} \left(\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} \Phi_{\vec{j}}^{(\boldsymbol{R})} - (2P_n(r_\rho))^{2\delta} \right) + \sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}} \sum_{\substack{\vec{j}\in T_{\vec{i}} \\ \mathbf{\Sigma}_{\vec{j}} \text{ not diagonal}}} \left(\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} \Phi_{\vec{j}}^{(\boldsymbol{R})} - (2P_n(r_\rho))^{2\delta} \right) \\
&\quad - \sum_{\vec{i}\in C_\delta^<} \sum_{\vec{j}\in C_\delta^<\setminus T_{\vec{i}}} (2P_n(r_\rho))^{2\delta} \qquad (3.136)
\end{aligned}
$$

where the last equality follows from $\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} = (2P_n(r_\rho))^\delta$ for $\vec{i} \in C_\delta^<$ such that $\Sigma_{\vec{i}}$ diagonal by Lemma 3.9.1. Then by (3.136),

$$
\begin{aligned}
&\sum_{\vec{i}\in C_\delta^<} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} W_{\vec{i}} - \left(\binom{p}{1} \binom{p-1}{\delta} (2P_n(r_\rho))^\delta \right)^2 \\
&\leq \left(\mu_{n,2\delta+2}(\mathbf{\Sigma}) - 1\right) (2P_n(r_\rho))^{2\delta} \left(\sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ not diagonal}}} \sum_{\vec{j}\in T_{\vec{i}}} 1 + \sum_{\substack{\vec{i}\in C_\delta^< \\ \mathbf{\Sigma}_{\vec{i}} \text{ diagonal}}} \sum_{\substack{\vec{j}\in T_{\vec{i}} \\ \mathbf{\Sigma}_{\vec{j}} \text{ not diagonal}}} 1 \right),
\end{aligned}
$$

where the inequality follows from Lemma 3.3.3 (c).

On the other hand, by (3.136),

$$\sum_{\vec{i}\in C_\delta^<} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} W_{\vec{i}} - \left(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\right)^2$$

$$\geq -(2P_n(r_\rho))^{2\delta}\left(\sum_{\substack{\vec{i}\in C_\delta^< \\ \boldsymbol{\Sigma}_{\vec{i}} \text{ not diagonal}}}\sum_{\vec{j}\in T_{\vec{i}}} 1 + \sum_{\substack{\vec{i}\in C_\delta^< \\ \boldsymbol{\Sigma}_{\vec{i}} \text{ diagonal}}}\sum_{\substack{\vec{j}\in T_{\vec{i}} \\ \boldsymbol{\Sigma}_{\vec{j}} \text{ not diagonal}}} 1\right) - \sum_{\vec{i}\in C_\delta^<}\sum_{\vec{j}\in C_\delta^<\setminus T_{\vec{i}}}(2P_n(r_\rho))^{2\delta}.$$

Combining th preceding two displays,

$$\left|\sum_{\vec{i}\in C_\delta^<} \mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} W_{\vec{i}} - \left(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\right)^2\right|$$

$$\leq \mu_{n,2\delta+2}(\boldsymbol{\Sigma})(2P_n(r_\rho))^{2\delta}\left(\sum_{\substack{\vec{i}\in C_\delta^< \\ \boldsymbol{\Sigma}_{\vec{i}} \text{ not diagonal}}}\sum_{\vec{j}\in T_{\vec{i}}} 1 + \sum_{\substack{\vec{i}\in C_\delta^< \\ \boldsymbol{\Sigma}_{\vec{i}} \text{ diagonal}}}\sum_{\substack{\vec{j}\in T_{\vec{i}} \\ \boldsymbol{\Sigma}_{\vec{j}} \text{ not diagonal}}} 1\right) + \sum_{\vec{i}\in C_\delta^<}\sum_{\vec{j}\in C_\delta^<\setminus T_{\vec{i}}}(2P_n(r_\rho))^{2\delta}$$

$$\leq \mu_{n,2\delta+2}(\boldsymbol{\Sigma})(2P_n(r_\rho))^{2\delta}\left(2\sum_{\substack{\vec{i}\in C_\delta^< \\ \boldsymbol{\Sigma}_{\vec{i}} \text{ not diagonal}}}\sum_{\vec{j}\in C_\delta^<} 1\right) + \sum_{\vec{i}\in C_\delta^<}\sum_{\vec{j}\in C_\delta^<\setminus T_{\vec{i}}}(2P_n(r_\rho))^{2\delta}$$

$$\overset{(*)}{\leq} \mu_{n,2\delta+2}(\boldsymbol{\Sigma})(2P_n(r_\rho))^{2\delta} 2\binom{p}{1}\binom{p-1}{\delta}\frac{(\delta+1)}{2((\delta-1)!)}p^\delta(\kappa-1) +$$

$$\binom{p}{1}\binom{p-1}{\delta}\frac{(\delta+1)^2}{\delta!}p^\delta\kappa(2P_n(r_\rho))^{2\delta}$$

$$\leq \frac{2(\delta+1)^2}{(\delta!)^2}(2p^{1+1/\delta}P_n(r_\rho))^{2\delta}\mu_{n,2\delta+2}(\boldsymbol{\Sigma})\frac{\kappa}{p}, \tag{3.137}$$

where step $(*)$ follows from Lemma 3.9.2 and (3.80).

Step 3:

Notice $\sum_{\vec{i}\in C_\delta^<}\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})} Z_{\vec{i}} = b_2$ as in (3.72) and thus satisfies the bound (3.87). Then by (3.133),

$$\left|\mathbb{E}\left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2 - \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\sum_{\ell=1}^{\delta+1}\ell\alpha(\ell,r_\rho) - \left(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\right)^2\right|$$

$$\leq \left|\sum_{\vec{i}\in C_\delta^<}\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\left(\Phi_{\vec{i}}^{(\boldsymbol{R})} + U_{\vec{i}}\right) - \binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\sum_{\ell=1}^{\delta+1}\ell\alpha(\ell,r_\rho)\right|$$

$$+\left|\sum_{\vec{i}\in C_\delta^<}\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}W_{\vec{i}}-\left(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta\right)^2\right|+b_2$$

$$\leq C_{n,\delta}\left((2p^{1+1/\delta}P_n(r_\rho))^\delta\left(1+2p^{1+1/\delta}P_n(r_\rho)\right)^\delta\mu_{n,2\delta+2}(\boldsymbol{\Sigma})\frac{\kappa}{p}+p(2pP_n(r_\rho))^{\delta+1}\left(1+2pP_n(r_\rho)\right)^\delta\right), \tag{3.138}$$

where the last inequality follows from (3.135), (3.137) and (3.87). The proof is then completed by $2p^{1+1/\delta}P_n(r_\rho)\leq\gamma$. □

3.12.2 Proof of Proposition 3.4.3

Proof of Proposition 3.4.3 (a):
By taking square of each terms in Lemma 3.3.8,

$$\left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2-2(\delta+1)N_{E_\delta}^{(\boldsymbol{R})}N_{E_{\delta+1}}^{(\boldsymbol{R})}\leq\left(N_{\check{V}_\delta}^{(\boldsymbol{R})}\right)^2\leq\left(N_{V_\delta}^{(\boldsymbol{R})}\right)^2\leq\left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2,$$

which then implies for $\bar{N}_\delta\in\left\{N_{\check{V}_\delta}^{(\boldsymbol{R})},N_{V_\delta}^{(\boldsymbol{R})}\right\}$

$$\left|\left(\bar{N}_\delta\right)^2-\left(N_{E_\delta}^{(\boldsymbol{R})}\right)^2\right|\leq 2(\delta+1)N_{E_\delta}^{(\boldsymbol{R})}N_{E_{\delta+1}}^{(\boldsymbol{R})}=2(\delta+1)\sum_{\vec{i}\in C_{\delta+1}^<}\sum_{\vec{j}\in C_\delta^<}\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}. \tag{3.139}$$

It suffices to establish an upper bound on $\mathbb{E}N_{E_{\delta+1}}^{(\boldsymbol{R})}N_{E_\delta}^{(\boldsymbol{R})}=\mathbb{E}\sum_{\vec{i}\in C_{\delta+1}^<}\sum_{\vec{j}\in C_\delta^<}\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}$.

Observe for $\vec{j}\in J_{\vec{i}}$,

$$\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}\leq\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\leq\mu_{n,\delta+2}(\boldsymbol{\Sigma}_{\vec{i}})(2P_n(r_\rho))^{\delta+1}.$$

For $\vec{j}\in T_{\vec{i}}$, $\left[\vec{j}\right]\cap\left[\vec{i}\right]=\emptyset$. Thus, if $\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}}$ is diagonal, $\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}=\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\mathbb{E}\Phi_{\vec{j}}^{(\boldsymbol{R})}=(2P_n(r_\rho))^{2\delta+1}$ by Lemma 3.9.1. Then for the general case that $\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}}$ is not necessarily diagonal, by Lemma 3.3.3 (c),

$$\mathbb{E}\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}\leq\mu_{n,2\delta+3}(\boldsymbol{\Sigma}_{\vec{i}\cup\vec{j}})(2P_n(r_\rho))^{2\delta+1}$$

It is straightforward by Lemma 3.3.3 (c), Lemma 3.9.1 and Lemma 3.9.4 that the conditions in Lemma 3.9.6 with $q=\delta+1$ and $\theta_{\vec{i},\vec{j}}=\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}$ are satisfied with $a=1$, $b=2P_n(2r_\rho)1(\delta\geq 2)+2P_n(r_\rho)1(\delta=1)$ and $z=2P_n(r_\rho)$. Moreover, $b/z\leq 2^{n-2}1(\delta\geq 2)+1$ by Lemma 3.14.1 (d).

Thus Lemma 3.9.7 with $q=\delta+1$ and $\theta_{\vec{i},\vec{j}}=\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}$, $a=1$, $b=2P_n(2r_\rho)1(\delta\geq 2)+2P_n(r_\rho)1(\delta=1)$ and $z=2P_n(r_\rho)$, together with the fact that $pz\leq p^{1+\frac{1}{\delta}}z=2p^{1+\frac{1}{\delta}}P_n(r_\rho)\leq\gamma$,

yield

$$\mathbb{E}N_{E_{\delta+1}}^{(\boldsymbol{R})}N_{E_\delta}^{(\boldsymbol{R})} = \mathbb{E}\sum_{\vec{i}\in C_{\delta+1}^<}\sum_{\vec{j}\in C_\delta^<}\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})} \le C_{n,\delta,\gamma}\left(1+\mu_{n,2\delta+3}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)p^{-1/\delta}. \tag{3.140}$$

The proof is then complete by the preceding display and (3.139).

□

We now present a few lemmas that are used in the proof of Proposition 3.4.3 (b) and (c). Recall $F_{ij}(r_\rho), H_{ij}(r_\rho), G_{ij}(r_\rho), F_{\vec{i}}(r_\rho)$ are defined in Section 3.10.3.

Lemma 3.12.1. *Suppose $p \ge n$. $1 \le \delta \le q \le p-1$. Then for any $\vec{i} \in C_q^<$, $\vec{j} \in C_\delta^<$, with probability* 1*,*

$$\left|\Phi_{\vec{i}}^{(\boldsymbol{P})}\Phi_{\vec{j}}^{(\boldsymbol{P})} - \Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}\right| \le \xi_{\vec{i},\vec{j}},$$

where

$$\xi_{\vec{i},\vec{j}} := 1\left(\bigcup_{m=1}^{q}\left(\left(H_{i_0i_m}\backslash G_{i_0i_m}\right)\bigcap H_{\vec{i},-m}\bigcap H_{\vec{j}}\right)\bigcup\bigcup_{\ell=1}^{\delta}\left(\left(H_{j_0j_\ell}\backslash G_{j_0j_\ell}\right)\bigcap H_{\vec{j},-\ell}\bigcap H_{\vec{i}}\right)\right).$$

Proof:

$$\left|\Phi_{\vec{i}}^{(\boldsymbol{P})}\Phi_{\vec{j}}^{(\boldsymbol{P})} - \Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}\right| = 1\left(\left(\bigcap_{m=1}^{q}F_{i_0i_m}\bigcap\bigcap_{\ell=1}^{\delta}F_{j_0j_\ell}\right)\triangle\left(\bigcap_{m=1}^{q}S_{i_0i_m}\bigcap\bigcap_{\ell=1}^{\delta}S_{j_0j_\ell}\right)\right) \le \xi_{\vec{i},\vec{j}}$$

where the inequality follows from (3.95) and Lemma 3.14.7 (a). □

Lemma 3.12.2. *Let $p \ge n \ge 4$, $1 \le \delta \le q \le p-1$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \le \frac{p}{2}$. Let t be any positive number, and suppose* (3.100) *holds. Then for any $\vec{i} \in C_q^<$, $\vec{j} \in C_\delta^<$, with probability* 1*,*

$$\xi_{\vec{i},\vec{j}}1\left(\mathcal{E}(t)\right) \le \eta_{\vec{i},\vec{j}}(t),$$

where

$$\begin{aligned}\eta_{\vec{i},\vec{j}}(t) :=& 1\left(\bigcup_{m=1}^{q}\left(\left(F_{i_0i_m}(\theta_1(t)r_\rho)\backslash F_{i_0i_m}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\bigcap F_{\vec{i},-m}(\theta_1(t)r_\rho)\bigcap F_{\vec{j}}(\theta_1(t)r_\rho)\right)\bigcup\right.\\ &\left.\bigcup_{\ell=1}^{\delta}\left(\left(F_{j_0j_\ell}(\theta_1(t)r_\rho)\backslash F_{j_0j_\ell}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\bigcap F_{\vec{j},-\ell}(\theta_1(t)r_\rho)\bigcap F_{\vec{i}}(\theta_1(t)r_\rho)\right)\right).\end{aligned}$$

Proof: By (3.101), $H_{ij}(r_\rho) \cap \mathcal{E}(t) \subset F_{ij}(\theta_1(t) r_\rho)$ and $G_{ij}(r_\rho) \cap \mathcal{E}(t) \supset F_{ij}\left(\frac{r_\rho}{\theta_1(t)}\right)$. Then

$$\xi_{\vec{i},\vec{j}} 1\left(\mathcal{E}(t)\right) \leq \eta_{\vec{i},\vec{j}}(t).$$

□

Lemma 3.12.3. *Let $p \geq n \geq 4$, $1 \leq \delta \leq q \leq p-1$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta)$. Suppose $\boldsymbol{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \leq \frac{p}{2}$. Let t be any positive number, and suppose* (3.100) *holds. Suppose additionally $2p^{1+\frac{1}{\delta}} P_n(r_\rho) \leq \gamma$. Then*

$$\mathbb{E} \sum_{\vec{i} \in C_q^<} \sum_{\vec{j} \in C_\delta^<} \eta_{\vec{i},\vec{j}}(t) \leq C_{n,q,\delta,\gamma} \left(\theta_1(t)\right)^{n(2\delta+q)} \left(1 + \mu_{n,q+\delta+2}(\boldsymbol{\Sigma}) \frac{\kappa - 1}{p}\right) \left(\sqrt{\frac{1}{p}} + \frac{t}{\sqrt{p}} + \frac{\tau}{p}\right) p^{1-\frac{q}{\delta}}.$$

Proof: Note

$$\sum_{\vec{i} \in C_q^<} \sum_{\vec{j} \in C_\delta^<} \eta_{\vec{i},\vec{j}}(t) = \left(\sum_{\vec{i} \in C_q^<} \sum_{\vec{j} \in J_{\vec{i}}} + \sum_{\vec{i} \in C_q^<} \sum_{\vec{j} \in T_{\vec{i}}} + \sum_{\vec{i} \in C_q^<} \sum_{\vec{j} \in N_{\vec{i}}}\right) \eta_{\vec{i},\vec{j}}(t).$$

Step 1: $\vec{j} \in T_{\vec{i}}$
By union bound for indicator function,

$$\eta_{\vec{i},\vec{j}}(t) \leq \eta_{\vec{i}}(t) 1\left(F_{\vec{j}}(\theta_1(t) r_\rho)\right) + \eta_{\vec{j}}(t) 1\left(F_{\vec{i}}(\theta_1(t) r_\rho)\right), \tag{3.141}$$

where $\eta_{\vec{i}}(t)$ is defined in (3.102) with δ replaced by q. Then for $\vec{j} \in T_{\vec{i}}$

$$\mathbb{E}\eta_{\vec{i},\vec{j}}(t) \leq \mathbb{E}\eta_{\vec{i}}(t) \mathbb{P}\left(F_{\vec{j}}(\theta_1(t) r_\rho)\right) + \mathbb{E}\eta_{\vec{j}}(t) \mathbb{P}\left(F_{\vec{i}}(\theta_1(t) r_\rho)\right). \tag{3.142}$$

Moreover, for $\vec{j} \in T_{\vec{i}}$, $\boldsymbol{\Sigma}_{\vec{i} \cup \vec{j}}$ is diagonal if and only if $\boldsymbol{\Sigma}_{\vec{i}}$ and $\boldsymbol{\Sigma}_{\vec{j}}$ are both diagonal.

Now suppose $\boldsymbol{\Sigma}_{\vec{i} \cup \vec{j}}$ is diagonal, by conditioning on $\boldsymbol{u}_{j_0}$

$$\mathbb{P}\left(F_{\vec{j}}(\theta_1(t) r_\rho)\right) = \left(2P_n(\theta_1(t) r_\rho)\right)^\delta, \quad \mathbb{P}\left(F_{\vec{i}}(\theta_1(t) r_\rho)\right) = \left(2P_n(\theta_1(t) r_\rho)\right)^q.$$

The preceding display, (3.142), Lemma 3.10.5 applying to $\mathbb{E}\eta_{\vec{j}}(t)$, and Lemma 3.10.5 with $\delta = q$ applying to $\mathbb{E}\eta_{\vec{i}}(t)$ yield

$$\mathbb{E}\eta_{\vec{i},\vec{j}}(t) \leq (\delta + q) 2 \left(P_n(r_\rho \theta_1(t)) - P_n\left(\frac{r_\rho}{\theta_1(t)}\right)\right) \left(2P_n\left(\theta_1(t) r_\rho\right)\right)^{q+\delta-1}.$$

For the general case that $\mathbf{\Sigma}_{\vec{i}\cup\vec{j}}$ is not necessarily diagonal, by Lemma 3.3.3 (c), for any $\vec{j} \in T_{\vec{i}}$

$$\mathbb{E}\eta_{\vec{i},\vec{j}}(t) \leq \mu_{n,q+\delta+2}(\mathbf{\Sigma}_{\vec{i}\cup\vec{j}})(\delta+q)2\left(P_n(r_\rho\theta_1(t)) - P_n\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\left(2P_n\left(\theta_1(t)r_\rho\right)\right)^{q+\delta-1}.$$

Then the condition in Lemma 3.9.5 (b) is satisfied with $\theta_{\vec{i},\vec{j}} = \eta_{\vec{i},\vec{j}}$, $z = 2P_n(r_\rho\theta_1(t))$ and $a = (\delta+q)\frac{P_n(r_\rho\theta_1(t)) - P_n\left(\frac{r_\rho}{\theta_1(t)}\right)}{P_n(r_\rho\theta_1(t))}$.

Step 2: $\vec{j} \in J_{\vec{i}}$
(3.141) implies

$$\eta_{\vec{i},\vec{j}}(t) \leq \eta_{\vec{i}}(t) + \sum_{\ell=1}^{\delta} 1\left(F_{j_0j_\ell}(\theta_1(t)r_\rho)\big\backslash F_{j_0j_\ell}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)1(F_{\vec{i}}(\theta_1(t)r_\rho)). \tag{3.143}$$

For $\vec{j} \in J_{\vec{i}}$, $j_0, j_\ell \in \left[\vec{i}\right]$. If $i_0 \in \{j_0, j_\ell\}$, without loss of generality, say $i_0 = j_\ell$ and $j_0 = i_\alpha$ for some $1 \leq \alpha \leq q$. Then

$$\begin{aligned}
&\mathbb{E}1\left(F_{j_0j_\ell}(\theta_1(t)r_\rho)\big\backslash F_{j_0j_\ell}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)1(F_{\vec{i}}(\theta_1(t)r_\rho)) \\
=&\mathbb{P}\left(\left(F_{i_0i_\alpha}(\theta_1(t)r_\rho)\big\backslash F_{i_0i_\alpha}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\bigcap F_{\vec{i},-\alpha}(\theta_1(t)r_\rho)\right) \\
\leq&\mu_{n,q+1}(\mathbf{\Sigma}_{\vec{i}})2\left(P_n(r_\rho\theta_1(t)) - P_n\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\left(2P_n\left(\theta_1(t)r_\rho\right)\right)^{q-1},
\end{aligned} \tag{3.144}$$

where the last step follows from Lemma 3.10.5 with δ replace by q.

If $i_0 \notin \{j_0, j_\ell\}$, without loss of generality, say $j_0 = i_\alpha$, $j_\ell = i_\beta$ for some $1 \leq \alpha \neq \beta \leq q$. Suppose for now that $\mathbf{\Sigma}_{\vec{i}}$ is diagonal, then

$$\begin{aligned}
&\mathbb{E}1\left(F_{j_0j_\ell}(\theta_1(t)r_\rho)\big\backslash F_{j_0j_\ell}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)1(F_{\vec{i}}(\theta_1(t)r_\rho)) \\
=&\mathbb{E}1\left(F_{i_\alpha i_\beta}(\theta_1(t)r_\rho)\big\backslash F_{i_\alpha i_\beta}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)1(F_{\vec{i}}(\theta_1(t)r_\rho)) \\
\overset{(*)}{=}&(2P_n(\theta_1(t)r_\rho))^{q-2}\mathbb{E}1\left(F_{i_\alpha i_\beta}(\theta_1(t)r_\rho)\big\backslash F_{i_\alpha i_\beta}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)1(F_{i_0i_\alpha}(\theta_1(t)r_\rho))1(F_{i_0i_\beta}(\theta_1(t)r_\rho)) \\
\overset{(**)}{\leq}&2\left(P_n(\theta_1(t)r_\rho) - P_n\left(\frac{r_\rho}{\theta_1(t)}\right)\right)\left(2P_n(\theta_1(t)r_\rho)\right)^{q-1},
\end{aligned}$$

where the step $(*)$ follows from conditioning on $\boldsymbol{u}_{i_0}, \boldsymbol{u}_{i_\alpha}, \boldsymbol{u}_{i_\beta}$, and the step $(**)$ follows from dropping the term $1(F_{i_0i_\beta}(\theta_1(t)r_\rho))$ and then conditioning on $\boldsymbol{u}_{i_\alpha}$. Then for the general case that $\mathbf{\Sigma}_{\vec{i}}$

is not necessarily diagonal, by Lemma 3.3.3 (c),

$$\begin{aligned}&\mathbb{E}1\left(F_{j_0j_\ell}(\theta_1(t)r_\rho)\backslash F_{j_0j_\ell}\left(\frac{r_\rho}{\theta_1(t)}\right)\right)1(F_{\vec{i}}(\theta_1(t)r_\rho))\\ \leq&\mu_{n,q+1}(\mathbf{\Sigma}_{\vec{i}})2\left(P_n(\theta_1(t)r_\rho)-P_n\left(\frac{r_\rho}{\theta_1(t)}\right)\right)(2P_n(\theta_1(t)r_\rho))^{q-1}.\end{aligned}\tag{3.145}$$

By combining (3.143), (3.144), (3.145) and Lemma 3.10.5 with δ replace by q,

$$\mathbb{E}\eta_{\vec{i},\vec{j}}(t)\leq(q+\delta)\mu_{n,q+1}(\mathbf{\Sigma}_{\vec{i}})2\left(P_n(\theta_1(t)r_\rho)-P_n\left(\frac{r_\rho}{\theta_1(t)}\right)\right)(2P_n(\theta_1(t)r_\rho))^{q-1}.$$

Then the condition in Lemma 3.9.5 (a) with $\theta_{\vec{i},\vec{j}}=\eta_{\vec{i},\vec{j}}$, $z=2P_n(r_\rho\theta_1(t))$ and $a=(q+\delta)\frac{P_n(\theta_1(t)r_\rho)-P_n\left(\frac{r_\rho}{\theta_1(t)}\right)}{P_n(\theta_1(t)r_\rho)}$ is satisfied.

Step 3: $\vec{j}\in N_{\vec{i}}$

It is straightforward by Lemma 3.3.3 (c), Lemma 3.9.1 and Lemma 3.9.4 that the conditions in Lemma 3.9.6 with $\theta_{\vec{i},\vec{j}}=\eta_{\vec{i},\vec{j}}$ are satisfied with $a=a_1=(q+\delta)\frac{P_n(\theta_1(t)r_\rho)-P_n\left(\frac{r_\rho}{\theta_1(t)}\right)}{P_n(\theta_1(t)r_\rho)}$, $b=2P_n(2r_\rho\theta_1(t))1(\delta\geq 2)+2P_n(r_\rho\theta_1(t))1(\delta=1)$ and $z=2P_n(r_\rho\theta_1(t))$. Moreover, $b/z\leq 2^{n-2}1(\delta\geq 2)+1$ by Lemma 3.14.1 (d).

Thus by Lemma 3.9.7 with $\theta_{\vec{i},\vec{j}}=\eta_{\vec{i},\vec{j}}$, $a=(q+\delta)\frac{P_n(\theta_1(t)r_\rho)-P_n\left(\frac{r_\rho}{\theta_1(t)}\right)}{P_n(\theta_1(t)r_\rho)}$, $b=2P_n(2r_\rho\theta_1(t))1(\delta\geq 2)+2P_n(r_\rho\theta_1(t))1(\delta=1)$ and $z=2P_n(r_\rho\theta_1(t))$

$$\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in C_\delta^<}\eta_{\vec{i},\vec{j}}\leq C_{n,q,\delta}\left(p^{1+\frac{1}{\delta}}z\right)^q\left(1+(p^{1+\frac{1}{\delta}}z)^\delta\right)(1+pz)^{\delta-1}\left(1+\mu_{n,q+\delta+2}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right)ap^{1-\frac{q}{p}}.\tag{3.146}$$

Step 4

Observe

$$pz\leq p^{1+\frac{1}{\delta}}z\leq(\theta_1(t))^{n-2}\,2p^{1+\frac{1}{\delta}}P_n(r_\rho)\leq(\theta_1(t))^{n-2}\,\gamma\tag{3.147}$$

where the second inequality follows from Lemma 3.14.1 (d). Moreover by Lemma 3.14.1 (c) and the fact that $\theta_1(t)\geq 1$,

$$a\leq(q+\delta)\frac{P_n(\theta_1(t)r_\rho)-P_n\left(\frac{r_\rho}{\theta_1(t)}\right)}{P_n(r_\rho)}\leq(q+\delta)(n-2)\,(\theta_1(t))^{n-3}\left(\theta_1(t)-\frac{1}{\theta_1(t)}\right).\tag{3.148}$$

Plug (3.147) and (3.148) into (3.146) and by the fact that $\theta_1(t) \geq 1$,

$$\begin{aligned}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in C_\delta^<}\eta_{\vec{i},\vec{j}} \leq & C_{n,q,\delta,\gamma}\left(\theta_1(t)\right)^{n(2\delta+q)}\left(1+\mu_{n,q+\delta+2}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right)\left(\theta_1(t)-\frac{1}{\theta_1(t)}\right)p^{1-\frac{q}{p}} \\ \leq & C_{n,q,\delta,\gamma}\left(\theta_1(t)\right)^{n(2\delta+q)}\left(1+\mu_{n,q+\delta+2}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right)\left(\sqrt{\frac{n}{p}}+\frac{t}{\sqrt{p}}+n\frac{\tau}{p}\right)p^{1-\frac{q}{p}},\end{aligned}$$

where the last inequality follows from Lemma 3.14.3 (d) and (3.101). □

Lemma 3.12.4. *Let $p \geq n \geq 4$ and $\boldsymbol{X} \sim \mathcal{VE}(\boldsymbol{\mu}, \mathbf{\Sigma}, \theta)$. Suppose $\mathbf{\Sigma}$, after some row-column permutation, is (τ, κ) sparse with $\tau \leq \frac{p}{2}$. Consider $1 \leq \delta \leq p-2$ and let $q \in \{\delta, \delta+1\}$. Suppose $2p^{1+\frac{1}{\delta}}P_n(r_\rho) \leq \gamma$ and $\left(\sqrt{\frac{n-1}{p}}+\sqrt{\frac{\delta \ln p}{p}}\right) \leq c$ hold for some positive and small enough universal constant c. Then*

$$\mathbb{E}\left|N_{E_q}^{(\boldsymbol{P})}N_{E_\delta}^{(\boldsymbol{P})}-N_{E_q}^{(\boldsymbol{R})}N_{E_\delta}^{(\boldsymbol{R})}\right| \leq C_{n,\delta,\gamma}\left(1+\mu_{n,q+\delta+2}(\mathbf{\Sigma})\frac{\kappa-1}{p}\right)\left(\frac{\sqrt{\ln p}}{\sqrt{p}}+\frac{\tau}{p}\right)p^{1-\frac{q}{\delta}}.$$

Proof: For $k \in \{\boldsymbol{R}, \boldsymbol{P}\}$,

$$N_{E_q}^{(k)}N_{E_\delta}^{(k)} = \sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in C_\delta^<}\Phi_{\vec{i}}^{(k)}\Phi_{\vec{j}}^{(k)}.$$

Thus

$$\begin{aligned}&\mathbb{E}\left|N_{E_q}^{(\boldsymbol{P})}N_{E_\delta}^{(\boldsymbol{P})}-N_{E_q}^{(\boldsymbol{R})}N_{E_\delta}^{(\boldsymbol{R})}\right| \\ \leq & \mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in C_\delta^<}\left|\Phi_{\vec{i}}^{(\boldsymbol{P})}\Phi_{\vec{j}}^{(\boldsymbol{P})}-\Phi_{\vec{i}}^{(\boldsymbol{R})}\Phi_{\vec{j}}^{(\boldsymbol{R})}\right|1\left(\mathcal{E}(t)\right)+\binom{p}{1}\binom{p-1}{q}\binom{p}{1}\binom{p-1}{\delta}\mathbb{P}(\mathcal{E}^c(t)) \\ \leq & \mathbb{E}\sum_{\vec{i}\in C_q^<}\sum_{\vec{j}\in C_\delta^<}\eta_{\vec{i},\vec{j}}+\frac{p^{q+\delta+2}}{\delta!q!}2\exp(-c_1t^2) \qquad (3.149)\end{aligned}$$

where the first inequality follows from $0 \leq N_{E_\delta}^{(k)} \leq \binom{p}{1}\binom{p-1}{\delta}$ for both $k = \boldsymbol{R}$ and $k = \boldsymbol{P}$, and the second inequality follows from Lemma 3.12.1, Lemma 3.12.2 and (3.97).

Choose $t = s_0\sqrt{\ln p}$ with $s_0 = \sqrt{\left(\frac{9}{2}+2\delta\right)/c_1}$. Since $q \in \{\delta, \delta+1\}$, $s_0 \geq \sqrt{\left(\frac{3}{2}+q+\delta+\frac{q}{\delta}\right)/c_1}$. Then

$$2\exp(-c_1t^2) \leq 2\exp\left(-\left(\frac{3}{2}+q+\delta+\frac{q}{\delta}\right)\ln p\right) = \frac{2}{p^{\frac{3}{2}+q+\delta+\frac{q}{\delta}}}.$$

Moreover, for any $c < \frac{1}{2\max\left\{\sqrt{\left(\frac{9}{2}+2\delta\right)/c_1},1\right\}\sqrt{2}C_1}$,

$$\left(\sqrt{\frac{n-1}{p}}+\sqrt{\frac{\delta\ln p}{p}}\right) \leq c$$

implies

$$\sqrt{2}C_1\left(\sqrt{\frac{n-1}{p}}+s_0\sqrt{\frac{\ln p}{p}}\right)\leq\frac{1}{2}, \tag{3.150}$$

which is (3.100) with $t=s_0\sqrt{\ln p}$. Then apply Lemma 3.12.3 with $t=s_0\sqrt{\ln p}$ to (3.149),

$$\begin{aligned}
&\mathbb{E}\left|N_{E_q}^{(\boldsymbol{P})}N_{E_\delta}^{(\boldsymbol{P})}-N_{E_q}^{(\boldsymbol{R})}N_{E_\delta}^{(\boldsymbol{R})}\right| \\
\leq& C_{n,q,\delta,\gamma}\left(\theta_1(s_0\ln p)\right)^{n(2\delta+q)}\left(1+\mu_{n,q+\delta+2}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)\left(\sqrt{\frac{1}{p}}+\frac{s_0\sqrt{\ln p}}{\sqrt{p}}+\frac{\tau}{p}\right)p^{1-\frac{q}{\delta}}+ \\
&\frac{2}{\delta!q!\sqrt{p}}p^{1-\frac{q}{\delta}} \\
\leq& C_{n,q,\delta,\gamma}\left(1+\mu_{n,q+\delta+2}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)\left(s_0\frac{\sqrt{\ln p}}{\sqrt{p}}+\frac{\tau}{p}\right)p^{1-\frac{q}{\delta}}+\frac{2}{\delta!q!\sqrt{p}}p^{1-\frac{q}{\delta}} \\
\leq& C_{n,q,\delta,\gamma}\left(1+\mu_{n,q+\delta+2}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)\left(\frac{\sqrt{\ln p}}{\sqrt{p}}+\frac{\tau}{p}\right)p^{1-\frac{q}{\delta}} \\
\leq& C_{n,\delta,\gamma}\left(1+\mu_{n,q+\delta+2}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)\left(\frac{\sqrt{\ln p}}{\sqrt{p}}+\frac{\tau}{p}\right)p^{1-\frac{q}{\delta}},
\end{aligned}$$

where the second inequality follows from $\theta_1\left(s_0\sqrt{\ln p}\right)\leq 9+4(n-1)=4n+5$ by (3.150) and $\tau\leq p/2$; and the last step follows from $q\in\{\delta,\delta+1\}$. □

Proof of Proposition 3.4.3 (b) and (c):

(b) It follows directly from Lemma 3.12.4 with $q=\delta$.

(c) By taking square of each terms in Lemma 3.3.8,

$$\left(N_{E_\delta}^{(\boldsymbol{P})}\right)^2-2(\delta+1)N_{E_\delta}^{(\boldsymbol{P})}N_{E_{\delta+1}}^{(\boldsymbol{P})}\leq\left(N_{\check{V}_\delta}^{(\boldsymbol{P})}\right)^2\leq\left(N_{V_\delta}^{(\boldsymbol{P})}\right)^2\leq\left(N_{E_\delta}^{(\boldsymbol{P})}\right)^2,$$

which then implies for $\bar{N}_\delta\in\left\{N_{\check{V}_\delta}^{(\boldsymbol{P})},N_{V_\delta}^{(\boldsymbol{P})}\right\}$

$$\left|\bar{N}_\delta-\left(N_{E_\delta}^{(\boldsymbol{P})}\right)^2\right|\leq 2(\delta+1)\left(N_{E_{\delta+1}}^{(\boldsymbol{P})}N_{E_\delta}^{(\boldsymbol{P})}-N_{E_{\delta+1}}^{(\boldsymbol{R})}N_{E_\delta}^{(\boldsymbol{R})}\right)+2(\delta+1)N_{E_{\delta+1}}^{(\boldsymbol{R})}N_{E_\delta}^{(\boldsymbol{R})}.$$

By Lemma 3.12.4 with $q=\delta+1$,

$$\mathbb{E}\left|N_{E_{\delta+1}}^{(\boldsymbol{P})}N_{E_\delta}^{(\boldsymbol{P})}-N_{E_{\delta+1}}^{(\boldsymbol{R})}N_{E_\delta}^{(\boldsymbol{R})}\right|\leq C_{n,\delta,\gamma}\left(1+\mu_{n,2\delta+3}(\boldsymbol{\Sigma})\frac{\kappa-1}{p}\right)\left(\frac{\sqrt{\ln p}}{\sqrt{p}}+\frac{\tau}{p}\right)p^{-\frac{1}{\delta}}.$$

The proof is then completed by combining the preceding two displays, (3.140) and the fact that

$$\left(\frac{\sqrt{\ln p}}{\sqrt{p}} + \frac{\tau}{p}\right) \leq 1.$$

□

3.13 Proofs in Section 3.5

3.13.1 Proofs of Lemma 3.5.3, Lemma 3.5.4 and Corollary 3.5.5

Proof of Lemma 3.5.3: When $\delta = 2$, $\alpha_2 = 0$ since either both vertices have the maximum degree 1 or none. Moreover,

$$\begin{aligned}
\alpha_3 =& \mathbb{P}(\|\tilde{\boldsymbol{u}}_1 - \tilde{\boldsymbol{u}}_2\|_2 \leq 1) \\
=& \mathbb{E}\mathbb{P}(\|\tilde{\boldsymbol{u}}_1 - \tilde{\boldsymbol{u}}_2\|_2 \leq 1 | \tilde{\boldsymbol{u}}_1) \\
\overset{(*)}{=}& \mathbb{E}\frac{1}{\text{Vol}(B^{n-2})} \times 2 \times \frac{\pi^{(n-3)/2}}{\Gamma(\frac{n-3}{2}+1)} \int_0^{\arccos(\frac{\|\tilde{\boldsymbol{u}}_1\|_2}{2})} \sin^{n-2}(\theta) d\theta \\
\overset{(**)}{=}& \frac{1}{\text{Vol}(B^{n-2})} \times 2 \times \frac{\pi^{(n-3)/2}}{\Gamma(\frac{n-3}{2}+1)} \frac{\text{Area}(S^{n-3})}{\text{Vol}(B^{n-2})} \int_0^1 r^{n-3} \int_0^{\arccos(\frac{r}{2})} \sin^{n-2}(\theta) d\theta dr \\
=& \frac{2(n-2)}{B(\frac{n-1}{2}, \frac{1}{2})} \int_0^1 r^{n-3} \int_0^{\arccos(\frac{r}{2})} \sin^{n-2}(\theta) d\theta dr
\end{aligned} \tag{3.151}$$

where step $(*)$ follows from the Subsection "Volume of a hyperspherical cap" of [Li11] and $\text{Vol}(B^{n-2})$ is the volume of B^{n-2}, step $(**)$ follows by observing the random quantity only depends on $\tilde{\boldsymbol{u}}_1$ through its Euclidean norm, and in the last step $B(\cdot, \cdot)$ is the Beta function. By Fubini's Theorem

$$\begin{aligned}
\int_0^1 r^{n-3} \int_0^{\arccos(\frac{r}{2})} \sin^{n-2}(\theta) d\theta dr =& \int_0^{\frac{\pi}{3}} \int_0^1 r^{n-3} \sin^{n-2}(\theta) dr d\theta + \int_{\frac{\pi}{3}}^{\frac{\pi}{2}} \int_0^{2\cos(\theta)} r^{n-3} \sin^{n-2}(\theta) dr d\theta \\
=& \frac{3}{2(n-2)} \int_0^{\frac{\pi}{3}} \sin^{n-2}(\theta) d\theta
\end{aligned}$$

Plug the preceding formula into (3.151), $\alpha_3 = \frac{3}{2} I_{\frac{3}{4}}(\frac{n-1}{2}, \frac{1}{2})$, where $I_x(a, b)$ is the regularized incomplete Beta function. $\alpha_1 = 1 - \alpha_3$ follows from $\alpha_2 = 0$. □

Proof of Lemma 3.5.4:

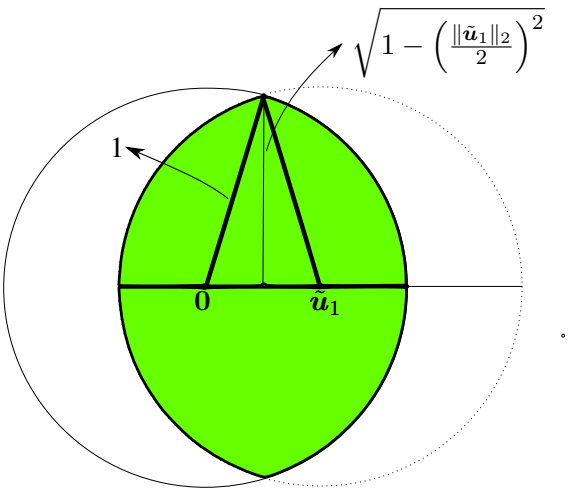


Figure 3.5: The solid circle represents the unit Euclidean ball B_2^n in $\mathbb{R}^n$ while the dash circle represents the unit ball centered at $\tilde{\boldsymbol{u}}_1$. Their intersection is the green region, which is contained in contained in the ball with center at $\tilde{\boldsymbol{u}}_1/2$ and with radius $\sqrt{1-\left(\frac{\|\tilde{\boldsymbol{u}}_1\|_2}{2}\right)^2}$.

a) Denote by $\deg(\cdot)$ the degree of a vertex in $\mathbf{Ge}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^m, 1; m, \mathscr{N}\right)$. Then by union bound,

$$\begin{aligned}\mathbb{P}\left(\mathbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; , n-2\right) \geq 1\right) \leq & \delta \mathbb{P}\left(\deg(\tilde{\boldsymbol{u}}_1) = \delta - 1\right) \\ = & \delta \mathbb{E}\left(\mathbb{P}(\deg(\tilde{\boldsymbol{u}}_1) = \delta - 1 | \tilde{\boldsymbol{u}}_1)\right) \\ = & \delta \mathbb{E}\left(\mathbb{P}\left(\|\tilde{\boldsymbol{u}}_1 - \tilde{\boldsymbol{u}}_2\|_2 \leq 1, \cdots, \|\tilde{\boldsymbol{u}}_1 - \tilde{\boldsymbol{u}}_\delta\|_2 \leq 1 | \, \tilde{\boldsymbol{u}}_1\right)\right) \\ = & \delta \mathbb{E}\left(\mathbb{P}\left(\|\tilde{\boldsymbol{u}}_1 - \tilde{\boldsymbol{u}}_2\|_2 \leq 1 | \, \tilde{\boldsymbol{u}}_1\right)^{\delta-1}\right), \end{aligned} \tag{3.152}$$

where the last equality follows by conditional independence.

As illustrated in Figure 3.5, $\mathbb{P}\left(\|\tilde{\boldsymbol{u}}_1 - \tilde{\boldsymbol{u}}_2\|_2 \leq 1 | \tilde{\boldsymbol{u}}_1\right)$ is the ratio between Lebesgue measure of green region and $|B_2^{n-2}|$. Moreover, the Lebesgue measure of the green region is less than $\left(1 - \left(\frac{\|\tilde{\boldsymbol{u}}_1\|_2}{2}\right)^2\right)^{\frac{n-2}{2}} |B_2^{n-2}|$. Then

$$\mathbb{P}\left(\|\tilde{\boldsymbol{u}}_1 - \tilde{\boldsymbol{u}}_2\|_2 \leq 1 | \tilde{\boldsymbol{u}}_1\right) \leq \left(1 - \left(\frac{\|\tilde{\boldsymbol{u}}_1\|_2}{2}\right)^2\right)^{\frac{n-2}{2}} \quad a.s. \tag{3.153}$$

By combining (3.152) and (3.153),

$$\mathbb{P}\left(\mathbf{NMD}\left(\{\tilde{\boldsymbol{u}}_i\}_{i=1}^{\delta}, 1; \delta, n-2\right) \geq 1\right) \leq \delta \mathbb{E}\left(1 - \left(\frac{\|\tilde{\boldsymbol{u}}_1\|_2}{2}\right)^2\right)^{\frac{(n-2)(\delta-1)}{2}}$$

$$
\begin{aligned}
&=\delta(n-2)\int_0^1\left(1-\frac{r^2}{4}\right)^{\frac{(n-2)(\delta-1)}{2}} r^{n-3}dr \qquad (3.154)\\
&=\delta(n-2)2^{n-3}B\left(\frac{1}{4};\frac{n-2}{2},\frac{(n-2)(\delta-1)}{2}+1\right),
\end{aligned}
$$

where the first equality follows from expressing the integral in polar coordinates, and the last step follows from changing the variables $r=2\sqrt{y}$.

b) Denote $f\left(r;\alpha,\beta\right)=\left(1-\frac{r^2}{4}\right)^{\alpha}r^{\beta}$. Then it is easy to verify that for any $\alpha,\beta>0$,

$$
\max_{r\in[0,1]} f\left(r;\alpha,\beta\right)=\begin{cases} f\left(1;\alpha,\beta\right)=\left(\frac{3}{4}\right)^{\alpha} & \text{if } 3\beta\geq 2\alpha,\\ f\left(\sqrt{\frac{4\beta}{2\alpha+\beta}};\alpha,\beta\right)=\left(\frac{2\alpha}{2\alpha+\beta}\right)^{\alpha}\left(\frac{4\beta}{2\alpha+\beta}\right)^{\frac{\beta}{2}} & \text{if } 3\beta\leq 2\alpha. \end{cases} \qquad (3.155)
$$

Moreover, $f(r;\alpha,\beta)$ is increasing on $[0,1]$ if $3\beta\geq 2\alpha$.

Let $\alpha=\frac{(n-2)(\delta-1)}{2}$ and $\beta=n-3$. If $\delta=2$, then for any $n\geq 4$, $3\beta\geq 2\alpha$ is satisfied. Then since $f(r;\alpha,\beta)$ is increasing on $[0,1]$,

$$
\int_0^1 f\left(r;\alpha,\beta\right)dr\leq\sqrt{\frac{4}{5}}f\left(\sqrt{\frac{4}{5}};\alpha,\beta\right)+\left(1-\sqrt{\frac{4}{5}}\right)f\left(1;\alpha,\beta\right). \qquad (3.156)
$$

If $\delta=3$, then for any $n\geq 5$, $3\beta\geq 2\alpha$ is satisfied and hence (3.156) holds. For $n=4$, $3\beta\leq 2\alpha$ is satisfied and by (3.155),

$$
\int_0^1 f\left(r;\alpha,\beta\right)dr\leq f\left(\sqrt{\frac{4\beta}{2\alpha+\beta}};\alpha,\beta\right)=f\left(\sqrt{\frac{4}{5}};\alpha,\beta\right). \qquad (3.157)
$$

If $\delta\geq 4$, it is easy to see for any $n\geq 4$, $3\beta\leq 2\alpha$ holds. By (3.155)

$$
\begin{aligned}
&\int_0^1 f\left(r;\alpha,\beta\right)dr\\
\leq& f\left(\sqrt{\frac{4\beta}{2\alpha+\beta}};\alpha,\beta\right)\\
=&\left(\frac{\delta-1}{\delta}\right)^{\frac{(n-2)(\delta-1)}{2}}\left(\frac{4}{\delta}\right)^{\frac{n-3}{2}}\left(\frac{n-2}{n-2-\frac{1}{\delta}}\right)^{\frac{\delta-1}{2}}\left(\left(\frac{n-2}{n-2-\frac{1}{\delta}}\right)^{\delta-1}\left(\frac{n-3}{n-2-\frac{1}{m}}\right)\right)^{\frac{n-3}{2}}\\
\leq&\exp\left(\frac{1}{4}\right)\left(\frac{\delta-1}{\delta}\right)^{\frac{(n-2)(\delta-1)}{2}}\left(\frac{4}{\delta}\right)^{\frac{n-3}{2}}, \qquad (3.158)
\end{aligned}
$$

where the last step follows from $\left(\frac{n-2}{n-2-\frac{1}{\delta}}\right)^{\delta-1}\left(\frac{n-3}{n-2-\frac{1}{\delta}}\right)\leq 1$ and $\left(\frac{n-2}{n-2-\frac{1}{\delta}}\right)^{\frac{\delta-1}{2}}\leq \exp\left(\frac{1}{4}\right)$.

Then (3.156), (3.157), (3.158) and the fact that $f\left(\sqrt{\frac{4}{5}};\alpha,\beta\right)=\left(\frac{4}{5}\right)^{\frac{(n-2)\delta-1}{2}}$ yields the conclusion.

□

Proof of Corollary 3.5.5: (a) By Lemma 3.5.4,

$$\begin{aligned}\sum_{\ell=2}^{\delta+1}\alpha_\ell \leq&\delta(n-2)2^{n-3}B\left(\frac{1}{4};\frac{n-2}{2},\frac{(n-2)(\delta-1)}{2}+1\right)\\ \leq&\begin{cases}\delta(n-2)\left(\left(\frac{4}{5}\right)^{\frac{(n-2)\delta-1}{2}}+\left(1-\sqrt{\frac{4}{5}}\right)\left(\frac{3}{4}\right)^{\frac{(n-2)(\delta-1)}{2}}\right) & \delta=2,3,\\ \delta(n-2)\exp\left(\frac{1}{4}\right)\left(\frac{\delta-1}{\delta}\right)^{\frac{(n-2)(\delta-1)}{2}}\left(\frac{4}{\delta}\right)^{\frac{n-3}{2}} & \delta\geq 4.\end{cases}\end{aligned}.$$

It suffices to prove $d_{\text{TV}}\left(\boldsymbol{\zeta}_{n,\delta},\delta_{\{1\}}\right)\leq\sum_{\ell=2}^{\delta+1}\alpha_\ell$. Notice that

$$\begin{aligned}d_{\text{TV}}\left(\boldsymbol{\zeta}_{n,\delta},\delta_{\{1\}}\right)=&\frac{1}{2}\sum_{\ell=1}^{\delta+1}\left|\boldsymbol{\zeta}_{n,\delta}(\ell)-\delta_{\{1\}}\left(\ell\right)\right|\\ =&\sum_{\ell=2}^{\delta+1}\boldsymbol{\zeta}_{n,\delta}(\ell) \qquad (3.159)\\ =&\frac{\sum_{\ell=2}^{\delta+1}\left(\alpha_\ell/\ell\right)}{\alpha_1+\sum_{\ell=2}^{\delta+1}\left(\alpha_\ell/\ell\right)}\\ \leq&\frac{\sum_{\ell=2}^{\delta+1}\alpha_\ell}{\alpha_1+\sum_{\ell=2}^{\delta+1}\alpha_\ell}\\ =&\sum_{\ell=2}^{\delta+1}\alpha_\ell. \qquad (3.160)\end{aligned}$$

(b) It follows from that

$$\left|\lambda_{n,\delta}(e_{n,\delta})-\frac{1}{\delta!}\left(e_n\right)^\delta\right|=\frac{1}{\delta!}\left(e_n\right)^\delta\left|\sum_{\ell=1}^{\delta+1}\left(\alpha_\ell/\ell\right)-1\right|\leq\frac{1}{\delta!}\left(e_n\right)^\delta\frac{3}{2}\sum_{\ell=2}^{\delta+1}\alpha_\ell\leq\frac{3}{2}\gamma_1\sum_{\ell=2}^{\delta+1}\alpha_\ell.$$

□

3.13.2 Proof of Lemma 3.5.6

Proof of Lemma 3.5.6: (a) Denote

$$I := \mathbb{P}\left(\mathbf{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1}, r; \delta+1, n-2\right) \geq 2 | \deg(\boldsymbol{u}_{\delta+1}') = \delta\right).$$

Then by union bound

$$\begin{aligned} I =& \mathbb{P}\left(\bigcup_{i=1}^{\delta} \deg(\boldsymbol{u}_i') = \delta | \deg(\boldsymbol{u}_{\delta+1}') = \delta\right) \\ \leq& \delta \mathbb{P}\left(\deg(\boldsymbol{u}_1') = \delta | \deg(\boldsymbol{u}_{\delta+1}') = \delta\right) \\ =& \delta \mathbb{P}\left(\deg(\boldsymbol{u}_1') = \delta, \deg(\boldsymbol{u}_{\delta+1}') = \delta\right) / \mathbb{P}\left(\deg(\boldsymbol{u}_{\delta+1}') = \delta\right). \end{aligned} \tag{3.161}$$

Notice that

$$\mathbb{P}\left(\deg(\boldsymbol{u}_{\delta+1}') = \delta\right) = \mathbb{E}\prod_{i=1}^{\delta} \mathbb{P}(\boldsymbol{u}_i' \in \mathrm{SC}(r, \boldsymbol{u}_{\delta+1}') | \boldsymbol{u}_{\delta+1}') = (P_n(r))^{\delta}, \tag{3.162}$$

where $\mathrm{SC}(r, \boldsymbol{u}_{\delta+1}')$ and $P_n(r)$ are defined in (3.27) and the paragraph after (3.27). Moreover

$$\begin{aligned} & \mathbb{P}\left(\deg(\boldsymbol{u}_1') = \delta, \deg(\boldsymbol{u}_{\delta+1}') = \delta\right) \\ =& \mathbb{E}\mathbb{P}\left(\deg(\boldsymbol{u}_1') = \delta, \deg(\boldsymbol{u}_{\delta+1}') = \delta | \boldsymbol{u}_1', \boldsymbol{u}_{\delta+1}'\right) \\ =& \mathbb{E}1(\|\boldsymbol{u}_1' - \boldsymbol{u}_{\delta+1}'\|_2 \leq r) \prod_{i=2}^{\delta} \mathbb{P}\left(\boldsymbol{u}_i' \in \mathrm{SC}(r, \boldsymbol{u}_1') \cap \mathrm{SC}(r, \boldsymbol{u}_{\delta+1}') | \boldsymbol{u}_1', \boldsymbol{u}_{\delta+1}'\right) \\ \leq& \mathbb{E}1(\|\boldsymbol{u}_1' - \boldsymbol{u}_{\delta+1}'\|_2 \leq r) \left(P_n(h(r, \|\boldsymbol{u}_1' - \boldsymbol{u}_{\delta+1}'\|_2))\right)^{\delta-1} \end{aligned} \tag{3.163}$$

where the last inequality follows from Lemma 3.13.1 with

$$h(r, d) = \sqrt{2 - \frac{2 - r^2}{\sqrt{1 - (\frac{d}{2})^2}}}.$$

Observing the random quantity in the expectation of (3.163) only depends the distance between $\|\boldsymbol{u}_1' - \boldsymbol{u}_{\delta+1}'\|_2$, replace $\boldsymbol{u}_{\delta+1}'$ with $\boldsymbol{v}_0 = (1, 0, \ldots, 0)$ will not change its value. Then

$$\mathbb{P}\left(\deg(\boldsymbol{u}_1') = \delta, \deg(\boldsymbol{u}_{\delta+1}') = \delta\right) \leq \mathbb{E}1(\|\boldsymbol{u}_1' - \boldsymbol{v}_0\|_2 \leq r) \left(P_n(h(r, \|\boldsymbol{u}_1' - \boldsymbol{v}_0\|_2))\right)^{\delta-1}. \tag{3.164}$$

Use the following coordinate system for each $\boldsymbol{u}_1' = (u_{j1} : 1 \le j \le n-1)^T$ in the region $\mathrm{SC}(r, \boldsymbol{v}_0)$:

$$\begin{cases} u_{11} = 1 - \frac{r^2 r_1^2}{2}, \\ u_{21} = r_1 r \sqrt{1 - \frac{r^2 r_1^2}{4}} \cos(\theta_2), \\ \vdots \\ u_{j1} = r_1 r \sqrt{1 - \frac{r^2 r_1^2}{4}} \cos(\theta_j) \prod_{m=2}^{j-1} \sin(\theta_m), \\ \vdots \\ u_{(n-2)1} = r_1 r \sqrt{1 - \frac{r^2 r_1^2}{4}} \sin(\theta_2) \cdots \sin(\theta_{n-3}) \cos(\theta_{n-2}), \\ u_{(n-1)1} = r_1 r \sqrt{1 - \frac{r^2 r_1^2}{4}} \sin(\theta_2) \cdots \sin(\theta_{n-3}) \sin(\theta_{n-2}), \end{cases}$$

where

$$r_1 \in [0,1], \theta_j \in [0,\pi] \text{ for } 2 \le j \le n-3 \text{ and } \theta_{n-2} \in [0, 2\pi). \tag{3.165}$$

Then the right hand side of (3.164) become

$$\begin{aligned} &\mathbb{E}1(\|\boldsymbol{u}_1' - \boldsymbol{v}_0\|_2 \le r) \left(P_n(h(r, \|\boldsymbol{u}_1' - \boldsymbol{v}_0\|_2))\right)^{\delta-1} \\ =&\frac{1}{\mathrm{Area}(S^{n-2})} \int_0^1 \left(P_n(h(r, r_1 r))\right)^{\delta-1} r^{n-2} r_1^{n-3} \left(1 - \frac{r^2 r_1^2}{4}\right)^{\frac{n-4}{2}} dr_1 \prod_{j=2}^{n-3} \int_0^\pi \sin^{n-2-j}(\theta_j) d\theta_j \\ =&\frac{1}{\int_0^\pi \sin^{n-3}(\theta) d\theta} \int_0^1 \left(P_n(h(r, r_1 r))\right)^{\delta-1} r^{n-2} r_1^{n-3} \left(1 - \frac{r^2 r_1^2}{4}\right)^{\frac{n-4}{2}} dr_1 \\ =&\frac{r^{n-2}}{B(\frac{n-2}{2}, \frac{1}{2})} \int_0^1 \left(P_n(h(r, r_1 r))\right)^{\delta-1} r_1^{n-3} \left(1 - \frac{r^2 r_1^2}{4}\right)^{\frac{n-4}{2}} dr_1 \\ \le&\frac{r^{n-2}}{B(\frac{n-2}{2}, \frac{1}{2})} \int_0^1 \left(P_n(h(r, r_1 r))\right)^{\delta-1} r_1^{n-3} dr_1. \end{aligned} \tag{3.166}$$

Plug (3.162), (3.164) and (3.166) into (3.161) and we obtain

$$\begin{aligned} I \le&\delta \frac{r^{n-2}}{B(\frac{n-2}{2}, \frac{1}{2}) P_n(r)} \int_0^1 \left(\frac{P_n(h(r, r_1 r))}{P_n(r)}\right)^{\delta-1} r_1^{n-3} dr_1 \\ =&\delta(n-2) \frac{c_n r^{n-2}}{P_n(r)} \int_0^1 \left(\frac{P_n(h(r, r_1 r))}{P_n(r)}\right)^{\delta-1} r_1^{n-3} dr_1 \end{aligned} \tag{3.167}$$

where the equality follows from $c_n = \frac{1}{(n-2)B(\frac{n-2}{2}, \frac{1}{2})}$. By Lemma 3.14.1 (a),

$$\frac{c_n r^{n-2}}{P_n(r)} \le \frac{1}{\left(1 - \frac{r^2}{4}\right)^{\frac{n-4}{2}}}. \tag{3.168}$$

Since when $0 < r_1 < 1$, $0 < h(r,d)/r < 1$, by Lemma 3.14.1 (e),

$$\begin{aligned}\frac{P_n(h(r,r_1r))}{P_n(r)} &\le \left(\frac{h(r,r_1r)}{r}\right)^{n-2}\left(\frac{1-\frac{h^2(r,r_1r)}{4}}{1-\frac{r^2}{4}}\right)^{\frac{n-4}{2}} \\ &\le \left(1-\left(\frac{r_1}{2}\right)^2\right)^{\frac{n-2}{2}}\frac{1}{\sqrt{1-\left(\frac{r_1r}{2}\right)^2}}\left(\frac{1-\frac{h^2(r,r_1r)}{4}}{(1-\frac{r^2}{4})\sqrt{1-\left(\frac{r_1r}{2}\right)^2}}\right)^{\frac{n-4}{2}} \end{aligned} \tag{3.169}$$

where the second inequality follows from

$$\left(\frac{h(r,r_1r)}{r}\right)^2 = \frac{1}{\sqrt{1-\left(\frac{r_1r}{2}\right)^2}}\left(\frac{-2(\frac{r_1}{2})^2}{1+\sqrt{1-(\frac{r_1r}{2})^2}}+1\right) \le \frac{1}{\sqrt{1-\left(\frac{r_1r}{2}\right)^2}}\left(-\left(\frac{r_1}{2}\right)^2+1\right).$$

Since $h^2(r,r_1r)$ is decreasing function of $r_1 \in [0,1]$, (3.169) become

$$\begin{aligned}\frac{P_n(h(r,r_1r))}{P_n(r)} &\le \left(1-\left(\frac{r_1}{2}\right)^2\right)^{\frac{n-2}{2}}\frac{1}{\sqrt{1-\left(\frac{r}{2}\right)^2}}\left(\frac{1-\frac{h^2(r,r)}{4}}{(1-\frac{r^2}{4})\sqrt{1-\left(\frac{r}{2}\right)^2}}\right)^{\frac{n-4}{2}} \\ &\le \left(1-\left(\frac{r_1}{2}\right)^2\right)^{\frac{n-2}{2}}\frac{1}{\sqrt{1-\left(\frac{r}{2}\right)^2}}\left(\frac{1}{(1-\frac{r^2}{4})}\right)^{\frac{n-4}{2}}, \end{aligned} \tag{3.170}$$

where the second inequality follows from

$$1-\frac{h^2(r,r)}{4} \le \sqrt{1-\frac{r^2}{4}}.$$

Plug (3.168) and (3.170) into (3.167),

$$\begin{aligned} I \le& \delta(n-2)\frac{1}{\left(1-\frac{r^2}{4}\right)^{\frac{n+\delta-5}{2}}}\left(\frac{1}{(1-\frac{r^2}{4})}\right)^{\frac{(n-4)(\delta-1)}{2}}\int_0^1\left(1-\left(\frac{r_1}{2}\right)^2\right)^{\frac{(n-2)(\delta-1)}{2}} r_1^{n-3}dr_1 \\ =& \bar{h}\left(\frac{1}{\sqrt{1-r^2/4}},n,\delta\right)\delta(n-2)\int_0^1\left(1-\left(\frac{r_1}{2}\right)^2\right)^{\frac{(n-2)(\delta-1)}{2}} r_1^{n-3}dr_1. \end{aligned}$$

(b) Since $\frac{1}{\sqrt{1-r^2/4}}$ is decreasing and $\bar{h}(x,n,\delta)$ as a function of x is increasing,

$$\bar{h}\left(\frac{1}{\sqrt{1-r^2/4}},n,\delta\right) \leq \begin{cases} \bar{h}\left(\left(\frac{5}{4}\right)^{\frac{1}{4}},n,\delta\right), & \delta=2,3 \\ \bar{h}\left(\left(\frac{\delta}{\delta-1}\right)^{\frac{1}{4}},n,\delta\right), & \delta\geq 4 \end{cases}$$
$$= \begin{cases} \left(\sqrt{\frac{5}{4}}\right)^{\frac{n+\delta-5}{2}}\left(\sqrt{\frac{5}{4}}\right)^{\frac{(n-2)(\delta-1)}{2}}, & \delta=2,3 \\ \left(\sqrt{\frac{\delta}{\delta-1}}\right)^{\frac{n+\delta-5}{2}}\left(\sqrt{\frac{\delta}{\delta-1}}\right)^{\frac{(n-2)(\delta-1)}{2}}, & \delta\geq 4 \end{cases}. \tag{3.171}$$

Then the proof is complete by combining part (a), Lemma 3.5.4 (b) and (3.171).

Similar to (3.160), we have

$$d_{\mathrm{TV}}(\boldsymbol{\zeta}_{n,\delta,\rho},\delta_{\{1\}}) \leq \sum_{\ell=2}^{\delta+1}\alpha(\ell,r_\rho) = \mathbb{P}\left(\mathbf{NMD}\left(\{\boldsymbol{u}_i'\}_{i=1}^{\delta+1},r_\rho;\delta+1,n-2\right)\geq 2|\deg(\boldsymbol{u}_{\delta+1}')=\delta\right).$$

where the equality follows from (3.44). Then the conclusion follows from part (a) and part (b) since r_ρ satisfies the condition there.

□

Lemma 3.13.1. *Let $n\geq 3$ and $0<r<\sqrt{2}$. If $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are two points in S^{n-2} with $\|\boldsymbol{z}_1-\boldsymbol{z}_2\|_2=d$ satisfy $2-2\sqrt{1-(d/2)^2}<r^2$, then*

$$\mathbb{P}\left(\boldsymbol{u}_1'\in SC(r,\boldsymbol{z}_1)\cap SC(r,\boldsymbol{z}_2)\right)\leq P_n(h(r,d))$$

where $\boldsymbol{u}_1'$ has distribution unif(S^{n-2}) and

$$h(r,d)=\sqrt{2-\frac{2-r^2}{\sqrt{1-(\frac{d}{2})^2}}}.$$

Proof: The proof is based on Figure 3.6 and we use $|\cdot|$ to represent the length of a line segment in this proof. In the right triangle $\mathbf{0}\boldsymbol{z}_3\boldsymbol{z}_1$, the line segment $\mathbf{0}\boldsymbol{z}_3$ has length $|\mathbf{0}\boldsymbol{z}_3|=\sqrt{1-(d/2)^2}$. In the right triangle $\boldsymbol{z}_1\boldsymbol{z}_3\boldsymbol{z}_5$, $|\boldsymbol{z}_3\boldsymbol{z}_5|=\sqrt{r^2-(d/2)^2}$. In the triangle $\mathbf{0}\boldsymbol{z}_3\boldsymbol{z}_5$, by law of Cosines,

$$\cos(\theta)=\frac{2-r^2}{2\sqrt{1-(\frac{d}{2})^2}}.$$

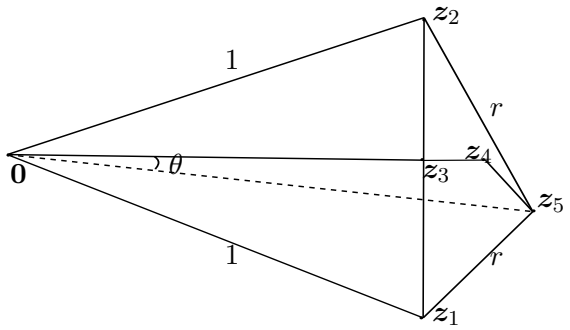


Figure 3.6: $\mathbf{0}$ is the origin in $\mathbb{R}^{n-2}$ and $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_4, \boldsymbol{z}_5$ are on S^{n-2}. $\boldsymbol{z}_3$ is the midpoint of z_1 and z_2, while z_4 is the midpoint of the shortest arc on S^{n-2} connecting $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. $\boldsymbol{z}_5$ is one of the two intersection points of the boundary $\mathrm{SC}(r, \boldsymbol{z}_1)$ and the boundary of $\mathrm{SC}(r, \boldsymbol{z}_2)$. The angle between line segment $\mathbf{0}\boldsymbol{z}_4$ and $\mathbf{0}\boldsymbol{z}_5$ is θ.

Then in the isosceles triangle, the line segment $\boldsymbol{z}_4\boldsymbol{z}_5$ has length

$$|\boldsymbol{z}_4\boldsymbol{z}_5| = 2\sin(\theta/2) = \sqrt{2(1-\cos(\theta))} = \sqrt{2 - \frac{2-r^2}{\sqrt{1-(\frac{d}{2})^2}}} = h(r,d).$$

It is easy to obtain $|\boldsymbol{z}_1\boldsymbol{z}_4| = \sqrt{2 - 2\sqrt{1-(d/2)^2}}$. The condition $2 - 2\sqrt{1-(d/2)^2} < r^2$ entails that $\mathrm{SC}(r, \boldsymbol{z}_1) \cap \mathrm{SC}(r, \boldsymbol{z}_2) \neq \emptyset$ and that $|\boldsymbol{z}_1\boldsymbol{z}_4| < |\boldsymbol{z}_4\boldsymbol{z}_5| = h(r,d)$. In this case $\mathrm{SC}(r, \boldsymbol{z}_1) \cap \mathrm{SC}(r, \boldsymbol{z}_2) \subset \mathrm{SC}(h(r,d), \boldsymbol{z}_4)$. Thus

$$\mathbb{P}\left(\boldsymbol{u}_1' \in \mathrm{SC}(r, \boldsymbol{z}_1) \cap \mathrm{SC}(r, \boldsymbol{z}_2)\right) \leq \mathbb{P}\left(\boldsymbol{u}_1' \in \mathrm{SC}(h(r,d), \boldsymbol{z}_4)\right) = P_n(h(r,d)).$$

□

3.14 Auxiliary lemmas

Lemma 3.14.1. *Let $P_n(r)$ be defined as in Section 3.3.3. Suppose $n \geq 4$.*

(a) Recall $c_n = \frac{b_n}{2(n-2)} = \frac{\Gamma((n-1)/2)}{(n-2)\sqrt{\pi}\Gamma((n-2)/2)} \leq 1$. Then

$$c_n r^{n-2}\left(1 - \frac{\min\{r^2, 4\}}{4}\right)^{\frac{n-4}{2}} \leq P_n(r) \leq c_n r^{n-2}$$

(b) $\lim_{r\to 0^+} P_n(r)/\left(c_n r^{n-2}\right) = 1$.

(c) Let $0 \leq \beta < 1 < \alpha$ and $0 < r \leq 2$. Then

$$P_n(\alpha r) - P_n(\beta r) \leq (n-2)P_n(r)\alpha^{n-3}(\alpha - \beta).$$

(d) Consider $\alpha > 1$ and $r > 0$. Then

$$P_n(\alpha r) \leq \alpha^{n-2} P_n(r).$$

(e) Consider $0 < \beta < 1$ and $0 < r < 2$. Then

$$P_n(\beta r) \leq \beta^{n-2} \left(\frac{1 - \frac{\beta^2 r^2}{4}}{1 - \frac{r^2}{4}} \right)^{\frac{n-4}{2}} P_n(r).$$

Proof: (a) It is easy to verify

$$P_n'(x) = \begin{cases} \frac{b_n}{2} x^{n-3} (1 - \frac{x^2}{4})^{\frac{n-4}{2}} & x < 2, \\ 0 & x \geq 2. \end{cases} \tag{3.172}$$

Consider $r > 0$. Then

$$\frac{P_n(r)}{c_n r^{n-2}} = \frac{P_n'(\xi)}{(n-2) c_n \xi^{n-3}} = \left(1 - \frac{\xi^2}{4} \right)^{\frac{n-4}{2}} \tag{3.173}$$

where in the first equality $\xi \in (0, \min\{r, 2\})$ due to Cauchy Mean Value Theorem and $P_n(r) \neq 0$, and the second equality follows from (3.172). (3.173) directly implies

$$\left(1 - \frac{(\min\{r, 2\})^2}{4} \right)^{\frac{n-4}{2}} \leq \frac{P_n(r)}{c_n r^{n-2}} \leq 1.$$

(b) It follows directly by taking limit $r \to 0^+$ in (3.173).

(c) Since $0 \leq \beta < 1 < \alpha$ and $0 < r \leq 2$, $P_n(\alpha r) - P_n(\beta r) > 0$ and $P_n(r) > 0$. Then

$$\begin{aligned} \frac{P_n(\alpha r) - P_n(\beta r)}{P_n(r)} &= \frac{(P_n(\alpha r) - P_n(\beta r)) - (P_n(\alpha \cdot 0) - P_n(\beta \cdot 0))}{P_n(r) - P_n(0)} \\ &= \frac{\frac{d}{dr} \left(P_n(\alpha r) - P_n(\beta r) \right) \big|_{r=\xi}}{\frac{d}{dr} P_n(r) \big|_{r=\xi}} \\ &= \frac{\alpha^{n-2} \left(1 - \frac{\alpha^2 \xi^2}{4} \right)^{\frac{n-4}{2}} - \beta^{n-2} \left(1 - \frac{\beta^2 \xi^2}{4} \right)^{\frac{n-4}{2}}}{\left(1 - \frac{\xi^2}{4} \right)^{\frac{n-4}{2}}} \\ &\leq \alpha^{n-2} - \beta^{n-2} \\ &\leq (n-2) \alpha^{n-3} (\alpha - \beta), \end{aligned} \tag{3.174}$$

where the second equality follows from Cauchy Mean Value Theorem with $\xi \in (0, r)$, the third equality follows from (3.172) together with the fact that the numerator has to be positive, which imply $\alpha\xi < 2$, the first inequality follows from $0 \leq \beta < 1 < \alpha$, and the last inequality follows from mean value theorem.

(d) When $r \geq 2$, $P_n(\alpha r) = P_n(r) = 1$ and the conclusion holds trivially. The case $0 < r < 2$ follows from (3.174) with $\beta = 0$.

(e) Consider $0 < \beta < 1$ and $0 < r < 2$. Then

$$\frac{P_n(\beta r)}{P_n(r)} = \frac{\beta^{n-2}\left(1 - \frac{\beta^2\xi^2}{4}\right)^{\frac{n-4}{2}}}{\left(1 - \frac{\xi^2}{4}\right)^{\frac{n-4}{2}}} \leq \frac{\beta^{n-2}\left(1 - \frac{\beta^2 r^2}{4}\right)^{\frac{n-4}{2}}}{\left(1 - \frac{r^2}{4}\right)^{\frac{n-4}{2}}},$$

where the equality follows from Cauchy Mean Value Theorem with $\xi \in (0, r)$.

□

Lemma 3.14.2. *Consider Z_1 and Z_2 be two discrete random variable support on $[\delta]$. Then*

$$d_{W}\left(\mathscr{L}(Z_1), \mathscr{L}(Z_2)\right) \leq \frac{\delta - 1}{2} \sum_{\ell=1}^{\delta} \left|\mathbb{P}(Z_1 = \ell) - \mathbb{P}(Z_2 = \ell)\right|.$$

Proof: By Remark 2.19 (iii) of Section 2.2 in [Vil03],

$$d_{\mathrm{W}}\left(\mathscr{L}(Z_1), \mathscr{L}(Z_2)\right) = \sum_{i=1}^{\delta-1} \left|\mathbb{P}(Z_1 \leq i) - \mathbb{P}(Z_2 \leq i)\right| \leq \sum_{i=1}^{\delta-1} \sum_{j=1}^{i} \left|\mathbb{P}(Z_1 = j) - \mathbb{P}(Z_2 = j)\right|.$$

On the other hand, from the above equality,

$$d_{\mathrm{W}}\left(\mathscr{L}(Z_1), \mathscr{L}(Z_2)\right) = \sum_{i=1}^{\delta-1} \left|\mathbb{P}(Z_1 \geq i+1) - \mathbb{P}(Z_2 \geq i+1)\right| \leq \sum_{i=1}^{\delta-1} \sum_{j=i+1}^{\delta} \left|\mathbb{P}(Z_1 = j) - \mathbb{P}(Z_2 = j)\right|.$$

Averaging the above two inequalities yields the desired conclusion. □

Lemma 3.14.3. *(a) Let p, p', m be positive integers such that $p \geq p'$. Then*

$$\prod_{i=0}^{m}(p - i) - \prod_{i=0}^{m}(p' - i) \leq (m + 1)\left(\prod_{i=0}^{m-1}(p - i)\right)(p - p').$$

(b) Let p, δ, κ be positive integers such that $\delta \le p-1$. Then

$$\prod_{\ell=1}^{\delta}(p-\ell) - \prod_{\ell=1}^{\delta}(p-\ell\kappa) \le \frac{\delta(\delta+1)}{2}(\kappa-1)\prod_{\ell=1}^{\delta-1}(p-\ell).$$

(c) $\left(\frac{1+x}{1-x}\right)^2$ is increasing function on $[0, \frac{1}{2}]$ and $\left(\frac{1+x}{1-x}\right)^2 \le 1+16x$ for $0 \le x \le \frac{1}{2}$.

(d) $1+x-\frac{1}{1+x} \le 2x$ for any $x \ge 0$.

Proof: (c) and (d) are simple quadratic inequalities and their proof are omitted.

(a) Let $f(x) = \prod_{i=0}^{m}(x-i)$. When $p' \ge m$, $f'(x) \le (m+1)\prod_{i=0}^{m-1}(p-i)$ and the conclusion then follows by mean value theorem. When $p' \le m-1$,

$$f(p) - f(p') \le f(p) \le (p-p')\prod_{i=0}^{m-1}(p-i).$$

(b) Let $f(x) = \prod_{\ell=1}^{\delta}(p-\ell x)$. When $p < \delta\kappa$,

$$f(1) - f(\kappa) \le f(1) \le (\delta\kappa-\delta)\prod_{\ell=1}^{\delta-1}(p-\ell) \le \frac{\delta(\delta+1)}{2}(\kappa-1)\prod_{\ell=1}^{\delta-1}(p-\ell).$$

When $p \ge \delta\kappa$, $f'(x) \ge -\frac{\delta(\delta+1)}{2}\prod_{\ell=1}^{\delta-1}(p-\ell)$ for $x \in [1,\kappa]$. Then the conclusion follows by mean value theorem.

□

Lemma 3.14.4. *For any integer-valued random variable Z_1 and Z_2,*

$$d_{TV}\left(\mathscr{L}(Z_1), \mathscr{L}(Z_2)\right) \le \mathbb{E}\left|Z_1 - Z_2\right|.$$

Proof:

$$\begin{aligned} d_{\text{TV}}\left(\mathscr{L}(Z_1), \mathscr{L}(Z_2)\right) &= \max_{A \text{ Borel measurable}} \left|\mathbb{P}(Z_1 \in A) - \mathbb{P}(Z_2 \in A)\right| \\ &= \max_{A \text{ Borel measurable}} \left|\mathbb{P}(Z_1 \in A, Z_1 \ne Z_2) - \mathbb{P}(Z_2 \in A, Z_1 \ne Z_2)\right| \\ &\le \max_{A \text{ Borel measurable}} \mathbb{P}(Z_1 \ne Z_2) \\ &= \mathbb{P}(|Z_1 - Z_2| \ge 1) \end{aligned}$$

$$\leq \mathbb{E}\left|Z_1 - Z_2\right|.$$

□

Lemma 3.14.5 (Perturbation Theory)**.** *Consider $\boldsymbol{D} \in \mathbb{S}^n$ and $\boldsymbol{E} \in \mathbb{S}^n$, where $\mathbb{S}^n$ is the set of all real symmetric matrices of dimension $n \times n$. Let $\{\lambda_i(\cdot)\}_{i=1}^n$ be the eigenvalues of corresponding matrix such that $\lambda_1(\cdot) \geq \lambda_2(\cdot) \geq \ldots \geq \lambda_n(\cdot)$.*

(a)

$$|\lambda_i(\boldsymbol{D}+\boldsymbol{E}) - \lambda_i(\boldsymbol{D})| \leq \|\boldsymbol{E}\|_2 \quad (i = 1, 2, \ldots, n)$$

(b) Assume $\boldsymbol{E} = \omega \boldsymbol{x}\boldsymbol{x}^T$, where $\boldsymbol{x} \in S^{n-1}$. If $\omega \geq 0$, then

$$\lambda_i(\boldsymbol{D}+\boldsymbol{E}) \in [\lambda_i(\boldsymbol{D}), \lambda_{i-1}(\boldsymbol{D})], \quad (i = 2, 3, \ldots, n),$$

while if $\omega \leq 0$, then

$$\lambda_i(\boldsymbol{D}+\boldsymbol{E}) \in [\lambda_{i+1}(\boldsymbol{D}), \lambda_i(\boldsymbol{D})], \quad (i = 1, 2, \ldots, n-1).$$

In either case, there exist nonnegative $m_1, m_2, \ldots, m_n$ such that

$$\lambda_i(\boldsymbol{D}+\boldsymbol{E}) = \lambda_i(\boldsymbol{D}) + m_i\omega, \quad (i = 1, 2, \ldots, n)$$

with $m_1 + m_2 + \cdots + m_n = 1$.

(c) Assume $\boldsymbol{E} = \sum_{i=1}^{m} \omega_i \boldsymbol{x}_i \boldsymbol{x}_i^T$, where $\{\boldsymbol{x}_i\}_{i=1}^m \subset S^{n-1}$ and $\omega_i \geq 0$ for all i. Then

$$\lambda_n(\boldsymbol{D}+\boldsymbol{E}) \geq \lambda_n(\boldsymbol{D}).$$

Proof: (a) and (b) is Corollary 8.1.6 and Theorem 8.1.8 in [GVL12]. (c) follows by induction on the smallest eigenvalue using part (b) for $\omega \geq 0$. □

Lemma 3.14.6. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2$ be two vectors on S^{n-1}, and $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ be an invertible matrix. Let $S_{min}(\boldsymbol{D})$ and $S_{max}(\boldsymbol{D})$ be respectively the largest and smallest singular value of $\boldsymbol{D}$. Define $\bar{\boldsymbol{z}}_i = \boldsymbol{D}\boldsymbol{x}_i$ and $\boldsymbol{z}_i = \bar{\boldsymbol{z}}_i / \|\bar{\boldsymbol{z}}_i\|_2$, $(i = 1, 2)$. Then,*

$$\frac{S_{min}(\boldsymbol{D})}{S_{max}(\boldsymbol{D})}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \leq \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 \leq \frac{S_{max}(\boldsymbol{D})}{S_{min}(\boldsymbol{D})}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$$

Proof: Part I (Upper Bound)

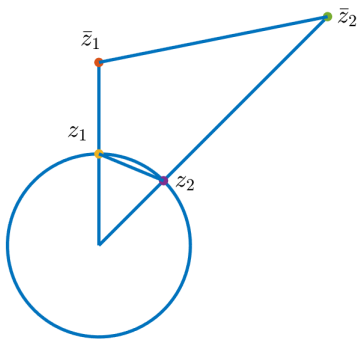


Figure 3.7: $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are the normalized vector of $\bar{\boldsymbol{z}}_1$ and $\bar{\boldsymbol{z}}_2$ respectively.

Denote $\angle(\cdot,\cdot)$ the angle between two vectors. By the Law of Cosines,

$$\cos(\angle(\boldsymbol{z}_1,\boldsymbol{z}_2)) = \frac{\|\boldsymbol{z}_1\|_2^2 + \|\boldsymbol{z}_2\|_2^2 - \|\boldsymbol{z}_1-\boldsymbol{z}_2\|_2^2}{2\times\|\boldsymbol{z}_1\|_2\times\|\boldsymbol{z}_2\|_2} = \frac{2-\|\boldsymbol{z}_1-\boldsymbol{z}_2\|_2^2}{2},$$

and

$$\cos(\angle(\bar{\boldsymbol{z}}_1,\bar{\boldsymbol{z}}_2)) = \frac{\|\bar{\boldsymbol{z}}_1\|_2^2 + \|\bar{\boldsymbol{z}}_2\|_2^2 - \|\bar{\boldsymbol{z}}_1-\bar{\boldsymbol{z}}_2\|_2^2}{2\times\|\bar{z}_1\|_2\times\|\bar{\boldsymbol{z}}_2\|_2}.$$

Observing $\angle(\boldsymbol{z}_1,\boldsymbol{z}_2) = \angle(\bar{\boldsymbol{z}}_1,\bar{\boldsymbol{z}}_2)$, right hand sides of the above two equations are equal. Solving for $\|\boldsymbol{z}_1-\boldsymbol{z}_2\|_2$, we get

$$\|\boldsymbol{z}_1-\boldsymbol{z}_2\|_2^2 = \frac{\|\bar{\boldsymbol{z}}_1-\bar{\boldsymbol{z}}_2\|_2^2}{\|\bar{\boldsymbol{z}}_1\|_2\|\bar{\boldsymbol{z}}_2\|_2} + \left(2-\frac{\|\bar{\boldsymbol{z}}_2\|_2}{\|\bar{\boldsymbol{z}}_1\|_2}-\frac{\|\bar{\boldsymbol{z}}_1\|_2}{\|\bar{\boldsymbol{z}}_2\|_2}\right) \le \frac{\|\bar{\boldsymbol{z}}_1-\bar{\boldsymbol{z}}_2\|_2^2}{\|\bar{\boldsymbol{z}}\|_2\|\bar{\boldsymbol{z}}_2\|_2}.$$

Therefore,

$$\begin{aligned}\|\boldsymbol{z}_1-\boldsymbol{z}_2\|_2 &\le \frac{\|\bar{\boldsymbol{z}}_1-\bar{\boldsymbol{z}}_2\|_2}{\sqrt{\|\bar{\boldsymbol{z}}_1\|_2\|\bar{\boldsymbol{z}}_2\|_2}}\\ &\le \frac{\mathrm{S}_{\max}(D)\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2}{\sqrt{\mathrm{S}_{\min}(\boldsymbol{D})\|\boldsymbol{x}_1\|_2\mathrm{S}_{\min}(\boldsymbol{D})\|\boldsymbol{x}_2\|_2}}\\ &= \frac{\mathrm{S}_{\max}(\boldsymbol{D})}{\mathrm{S}_{\min}(\boldsymbol{D})}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2.\end{aligned}$$

Part II(Lower Bound)

Define $\bar{\boldsymbol{x}}_i = \boldsymbol{D}^{-1}\boldsymbol{z}_i$, $(i=1,2)$. Notice for $\forall i\in\{1,2\}$, $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_i$ are parallel to each other, since $\boldsymbol{x}_i = \boldsymbol{D}^{-1}\bar{\boldsymbol{z}}_i$ and $\bar{\boldsymbol{z}}_i$ is parallel to $\boldsymbol{z}_i$. Thus, we conclude $\boldsymbol{x}_i = \bar{\boldsymbol{x}}_i/\|\bar{\boldsymbol{x}}_i\|_2$, $(i=1,2)$. Reversing the

role of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ in Part I, one has

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \leq \frac{\mathrm{S}_{\max}(\boldsymbol{D}^{-1})}{\mathrm{S}_{\min}(\boldsymbol{D}^{-1})}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2.$$

The lower bound follows from the relation $\frac{\mathrm{S}_{\max}(\boldsymbol{D}^{-1})}{\mathrm{S}_{\min}(\boldsymbol{D}^{-1})} = \frac{\mathrm{S}_{\max}(\boldsymbol{D})}{\mathrm{S}_{\min}(\boldsymbol{D})}$. □

Lemma 3.14.7. *Let $\{\mathcal{D}_i\}_{i=1}^m$, $\{\mathcal{F}_i\}_{i=1}^m$, $\{\mathcal{G}_i\}_{i=1}^m$ and $\{\mathcal{H}_i\}_{i=1}^m$ be sets satisfying*

$$\mathcal{G}_i \subset \mathcal{D}_i \subset \mathcal{H}_i, \quad \mathcal{G}_i \subset \mathcal{F}_i \subset \mathcal{H}_i, \quad (i = 1, 2, \ldots, m).$$

Then

(a)

$$\left(\bigcap_{i=1}^m \mathcal{D}_i\right) \triangle \left(\bigcap_{i=1}^m \mathcal{F}_i\right) \subset \bigcup_{i=1}^m (\mathcal{H}_i \backslash \mathcal{G}_i) \bigcap \left(\bigcap_{j=1}^m \mathcal{H}_j\right) = \bigcup_{i=1}^m \left((\mathcal{H}_i \backslash \mathcal{G}_i) \bigcap \left(\bigcap_{\substack{j=1 \\ j\neq i}}^m \mathcal{H}_j \right) \right).$$

(b)

$$\left(\bigcup_{i=1}^m \mathcal{D}_i\right) \triangle \left(\bigcup_{i=1}^m \mathcal{F}_i\right) \subset \bigcup_{i=1}^m (\mathcal{H}_i \backslash \mathcal{G}_i).$$

Proof: (a) Obviously,

$$\bigcap_{i=1}^m \mathcal{G}_i \subset \bigcap_{i=1}^m \mathcal{D}_i \subset \bigcap_{i=1}^m \mathcal{H}_i, \quad \bigcap_{i=1}^m \mathcal{G}_i \subset \bigcap_{i=1}^m \mathcal{F}_i \subset \bigcap_{i=1}^m \mathcal{H}_i. \tag{3.175}$$

Thus,

$$\left(\bigcap_{i=1}^m \mathcal{D}_i\right) \triangle \left(\bigcap_{i=1}^m \mathcal{F}_i\right) \subset \left(\bigcap_{i=1}^m \mathcal{H}_i\right) \setminus \left(\bigcap_{i=1}^m \mathcal{G}_i\right).$$

Take $\forall \omega \in (\bigcap_{i=1}^m \mathcal{H}_i) \setminus (\bigcap_{i=1}^m \mathcal{G}_i)$, we know $\omega \in \bigcap_{i=1}^m \mathcal{H}_i$ and $\omega \notin \bigcap_{i=1}^m \mathcal{G}_i$. The later fact shows $\exists j$ (which depends on ω) such that $\omega \notin \mathcal{G}_j$. Then,

$$\omega \in \left(\bigcap_{i=1}^m \mathcal{H}_i\right) \backslash \mathcal{G}_j \subset \mathcal{H}_j \backslash \mathcal{G}_j \subset \bigcup_{i=1}^m (\mathcal{H}_i \backslash \mathcal{G}_i). \tag{3.176}$$

The proof is completed by combining (3.175) and (3.176).

(b)

$$\begin{aligned}\left(\bigcup_{i=1}^{m}\mathcal{D}_i\right)\triangle\left(\bigcup_{i=1}^{m}\mathcal{F}_i\right) &= \left(\bigcup_{i=1}^{m}\mathcal{D}_i\right)^c\triangle\left(\bigcup_{i=1}^{m}\mathcal{F}_i\right)^c\\ &= \left(\bigcap_{i=1}^{m}\mathcal{D}_i^c\right)\triangle\left(\bigcap_{i=1}^{m}\mathcal{F}_i^c\right)\\ &\subset \bigcup_{i=1}^{m}(\mathcal{G}_i^c\backslash\mathcal{H}_i^c)\\ &= \bigcup_{i=1}^{m}(\mathcal{H}_i\backslash\mathcal{G}_i)\,,\end{aligned}$$

where the inclusion step follows from (a).

□

Lemma 3.14.8. *Let $\boldsymbol{Q}\in\mathbb{R}^{n\times m}(n\le m)$, with each column $\boldsymbol{q}_i$ being i.i.d. unif($\sqrt{n}S^{n-1}$). Let λ_{min} and λ_{max} be respectively the largest and smallest eigenvalue of $\frac{1}{m}\boldsymbol{Q}\boldsymbol{Q}^T$. Then with probability at least $1-2\exp(-ct^2)$,*

$$\left[1-C\left(\sqrt{\frac{n}{m}}+\frac{t}{\sqrt{m}}\right)\right]^2\le\lambda_{min}\le\lambda_{max}\le\left[1+C\left(\sqrt{\frac{n}{m}}+\frac{t}{\sqrt{m}}\right)\right]^2, \tag{3.177}$$

where c, C are absolute constants.

Proof: Let $\mathrm{S}_{\max}$, $\mathrm{S}_{\min}$ be respectively the largest and smallest singular value of $\boldsymbol{Q}$. Since $\boldsymbol{q}_i$ are isotropic random vector with subgaussian norm (or ψ_2 norm) being a constant, by applying Theorem 5.39 in [Ver12] to Q^T,

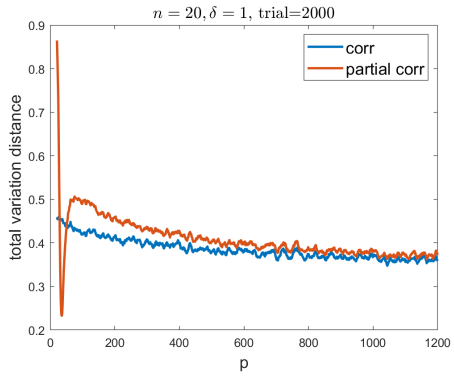
$$\sqrt{m}-C(\sqrt{n}+t)\le\mathrm{S}_{\min}\le\mathrm{S}_{\max}\le\sqrt{m}+C(\sqrt{n}+t), \tag{3.178}$$

holds with probability at least $1-2\exp(-ct^2)$, where c, C are absolute constants. The proof is completed by

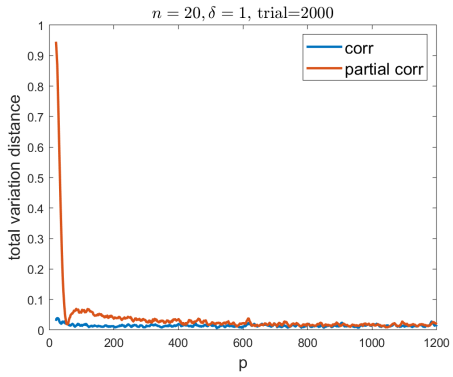
$$\lambda_{\max}=\frac{1}{m}\mathrm{S}_{\max}^2,\quad \lambda_{\min}=\frac{1}{m}\mathrm{S}_{\min}^2.$$

□

3.15 Numerical simulations and experiments

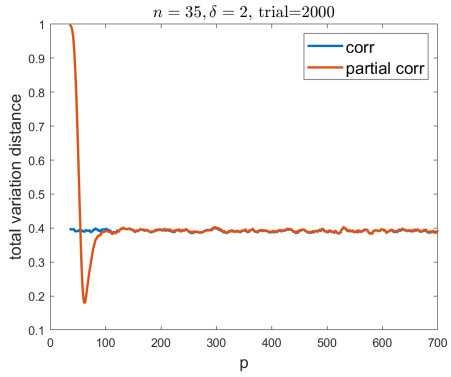


(a) limiting compound Poisson

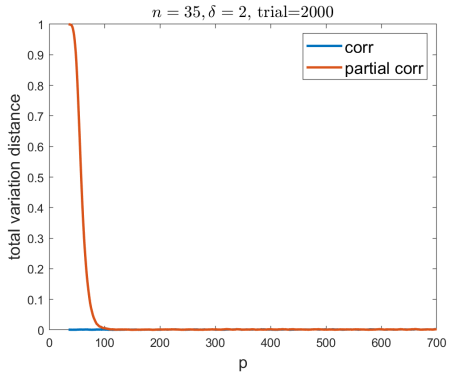


(b) finite p compound Poisson approximation

Figure 3.8: The vertical axis of (a) is $d_{\mathrm{TV}}(N_{V_1}^{(k)}, \mathrm{CP}(\lambda_{20,1}(1), \boldsymbol{\zeta}_{20,1}))$ as in Theorem 3.2.4 and that of (b) is $d_{\mathrm{TV}}(N_{V_1}^{(k)}, \mathrm{CP}(\lambda_{p,20,1,\rho}, \boldsymbol{\zeta}_{20,1,\rho}))$ as in Theorem 3.3.11. For both plots the samples are generated according to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being a $(\tau = p^{0.6}, \kappa = p^{0.8})$ sparse matrix for each p. The parameters are $n = 20$, $\delta = 1$ and the threshold ρ is chosen according to (3.12) with $e_{n,\delta} = 1$. The blue curve is for the empirical correlation graph (k=$\boldsymbol{R}$) and the red curve is for the empirical partial correlation graph (k=$\boldsymbol{P}$). Note since $\delta = 1$, $\boldsymbol{\zeta}_{20,1} = \delta_{\{2\}} = \boldsymbol{\zeta}_{20,1,\rho}$, by Example 3.5.1. As demonstrated by the plots, for both empirical correlation and partial correlation graphs, the total variations in (a) decrease very slowly while the total variations in (b) converge to 0 very fast, which has been analytically discussed in Remark 3.3.14.



(a) Poisson limit when $p \to \infty$



(b) Poisson approximation for finite p

Figure 3.9: The vertical axis of (a) is $d_{\text{TV}}(N^{(k)}_{V_\delta}, \text{Pois}(\frac{(e_{n,\delta})^\delta}{\delta!}))$, where we replaced $\text{CP}(\lambda_{n,\delta}(e_{n,\delta}), \boldsymbol{\zeta}_{n,\delta})$ in Theorem 3.2.4 by its approximation $\text{Pois}(\frac{(e_{n,\delta})^\delta}{\delta!})$ as discussed in Subsection 3.5.1. The vertical axis of (b) is $d_{\text{TV}}(N^{(k)}_{V_\delta}, \text{Pois}(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta))$, where we replaced $\text{CP}(\lambda_{p,n,\delta,\rho}, \boldsymbol{\zeta}_{n,\delta,\rho})$ in Theorem 3.3.11 by its approximation $\text{Pois}(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta)$ as discussed in Subsection 3.5.2. For both plots the samples are generated according to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being a $(\tau = p^{0.6}, \kappa = p^{0.8})$ sparse matrix for each p. The parameters are $n = 35$, $\delta = 2$ and the threshold ρ is chosen according to (3.12) with $e_{n,\delta} = 1$. Note the distributions of the increment $\boldsymbol{\zeta}_{35,2}$ in (a) and $\boldsymbol{\zeta}_{35,2,\rho}$ in (b) are both replaced by $\delta_{\{1\}}$ since $n = 35$ is sufficiently large for $\delta = 2$ as indicated by Figure 3.4 (b). That is, the number of samples $n = 35$ is large enough for Corollary 3.5.5 (a) and Lemma 3.5.6 (c) to be effective. The blue curve is for the empirical correlation graph (k=$\boldsymbol{R}$) and the red curve is for the empirical partial correlation graph (k=$\boldsymbol{P}$). As demonstrated by the plots, for both empirical correlation and partial correlation graphs, the total variations in (a) decrease very slowly while the total variations in (b) converge to 0 very fast. The fast convergence in Figure 3.9 (b) verifies the validity of using Poisson distribution $\text{Pois}(\binom{p}{1}\binom{p-1}{\delta}(2P_n(r_\rho))^\delta)$ to approximate the distribution of random quantities in $\{N^{(k)}_i : k = \boldsymbol{R}, i \in \{E_\delta, \breve{V}_\delta, V_\delta\}\}$ for large n. The extremely slow decrease in Figure 3.9 (a) is due to the slow convergence of Theorem 3.2.4, which has been extensively discussed in Remark 3.3.14. This specific example indicates the slow convergence of Theorem 3.2.4 is due to slow convergence of $\lambda_{p,n,\delta,\rho} \to \lambda_{n,\delta}$ since the distribution of increments in this large n case are both close to $\delta_{\{1\}}$.

CHAPTER 4

Future Directions

There are several future directions for follow-on work.

In Chapter 2, where the problem of parameter estimation in mixture of product distributions is discussed, the true number of components k is assumed to be known. This so-called "exact-fitted" assumption practically might not be available and it is of interest to consider the "over-fitted" scenario where only an upper bound on the true number of components is available. The parameter estimation in mixture models in "over-fitted" case but without product structure has been previously studied in [Ngu13, HN16b]. This generalization combines these existing work and the new product structure considered in this thesis and makes the study of mixture of product distribution more complete.

For the mixture of product distribution, a worthwhile direction is developing algorithms to efficiently estimate the parameters in the mixture of product distributions. The key insight is that the parameter information of the mixture of product distributions is contained in a mixture of corresponding Gaussian distribution by virtue of central limit theorem (density version), which is an important step of the proof in Chapter 2. Then based on this observation, and some recent developments in efficiently estimating parameters of location mixture of Gaussian [DWYZ20], it is worthwhile to explore some moment based algorithms to efficiently estimate the parameters in the mixture of product distributions.

For screening in high dimensional data in Chapter 3, a future direction is to extend the results in Chapter 3 to the setting of nonparametric data or functional data. Note the regime that the number of variables increases to infinity while the number of samples is fixed is indeed a fixed number of samples of (countably) infinite-dimensional data. This view explains the motive to extend such results to more general infinite-dimensional setup like functional data. This line of extension involves defining appropriate sparsity structure in the setting of nonparametric data or functional data since the results in Chapter 3 are built on the sparsely correlated assumption.

It is also important to study the statistical and computational problems involving both heterogeneity and high dimensions. One example of such a problem is the mixture of location Gaussian with the unknown location parameters that span an unknown low dimension space in the high

dimension where the data lie in. Intuitively, the geometric properties of concentration around the spherical shell for high dimensional gaussian and the rotation invariance for the simple case where the covariance matrix is identity will play crucial roles. Such a problem has wide applicability but seems open for theoretical study.

BIBLIOGRAPHY

[AF90] Theodore W Anderson and Kai-Tai Fang. Theory and applications of elliptically contoured and related distributions. Technical report, STANFORD UNIV CA DEPT OF STATISTICS, 1990.

[AK06] Charalambos D. Aliprantis and Border C. Kim. *Infinite dimensional analysis: A Hitchhiker's Guide*. Springer-Verlag Berlin Heidelberg, third edition, 2006.

[Ald85] David J. Aldous. Exchangeability and related topics. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg.

[AMR09] Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

[And92] Theodore W Anderson. Nonnormal multivariate distributions: Inference based on elliptically contoured distributions. Technical report, STANFORD UNIV CA DEPT OF STATISTICS, 1992.

[And03] Theodore W Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, third edition, 2003.

[Ant74] Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 11 1974.

[Bar01] Andrew D Barbour. Topics in poisson approximation. *Handbook of Statistics*, 19:79–115, 2001.

[BGG+07] Luca Brandolini, Giacomo Gigante, Allan Greenleaf, Alexander Iosevich, Andreas Seeger, and Giancarlo Travaglini. Average decay estimates for fourier transforms of measures supported on curves. *The Journal of geometric analysis*, 17(1):15–40, 2007.

[Bil96] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, third edition, 1996.

[CF06] Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM computing surveys (CSUR)*, 38(1):2–es, 2006.

[Che95] Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1):221–233, 02 1995.

[Cin11] Erhan Cinlar. *Probability and stochastics*, volume 261 of *Graduate Texts in Mathematics*. Springer Science & Business Media, 2011.

[CKG19] Xuan Cao, Kshitij Khare, and Malay Ghosh. Posterior graph selection and estimation consistency for high-dimensional bayesian dag models. *The Annals of Statistics*, 47(1):319–348, 2019.

[CLOP19] Federico Camerlenghi, Antonio Lijoi, Peter Orbanz, and Igor Prünster. Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92, 2019.

[Dav77] Philip David. Spherical matrix distributions and a multivariate model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):254–261, 1977.

[DR17] Onkar Dalal and Bala Rajaratnam. Sparse gaussian graphical model estimation via alternating minimization. *Biometrika*, 104(2):379–395, 2017.

[DWYZ20] Natalie Doss, Yihong Wu, Pengkun Yang, and Harrison H Zhou. Optimal estimation of high-dimensional gaussian mixtures. *arXiv preprint arXiv:2002.05818*, 2020.

[EHN05] Ryan Elmore, Peter Hall, and Amnon Neeman. An application of classical invariant theory to identifiability in nonparametric mixtures. In *Annales de l'institut Fourier*, volume 55, pages 1–28, 2005.

[Fel08] Willliam Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, third edition, 2008.

[Fer73] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 03 1973.

[GVDV01] Subhashis Ghosal and Aad W Van Der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263, 2001.

[GvdV17] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2017.

[GVL12] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3 of *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins University Press, fourth edition, 2012.

[HHMW10] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian Nonparametrics*, volume 28 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2010.

[HK18] Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018.

[HN16a] Nhat Ho and XuanLong Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726–2755, 2016.

[HN16b] Nhat Ho and XuanLong Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307, 2016.

[HN19] Nhat Ho and XuanLong Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *SIAM Journal on Mathematics of Data Science*, 1(4):730–758, 2019.

[HNPE05] Peter Hall, Amnon Neeman, Reza Pakyari, and Ryan Elmore. Nonparametric inference in multivariate mixtures. *Biometrika*, 92(3):667–678, 2005.

[HR11] Alfred Hero and Bala Rajaratnam. Large-scale correlation screening. *Journal of the American Statistical Association*, 106(496):1540–1552, 2011.

[HR12] Alfred Hero and Bala Rajaratnam. Hub discovery in partial correlation graphs. *IEEE Transactions on Information Theory*, 58(9):6064–6078, 2012.

[HR15a] Alfred Hero and Bala Rajaratnam. Foundational principles for large-scale inference: Illustrations through correlation mining. *Proceedings of the IEEE*, 104(1):93–110, 2015.

[HR15b] Alfred Hero and Bala Rajaratnam. Large-scale correlation mining for biomolecular network discovery. Technical report, STANFORD UNIV CA DEPT OF STATISTICS, 2015.

[HT00] TP Hettmansperger and Hoben Thomas. Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):811–825, 2000.

[HZ03] Peter Hall and Xiao-Hua Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics*, 31(1):201–224, 2003.

[Kal06] Olav Kallenberg. *Probabilistic symmetries and invariance principles*. Probability and Its Applications. Springer Science & Business Media, 2006.

[KC14] Eric D Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*, volume 65 of *Use R!*. Springer, 2014.

[Kol09] Eric D Kolaczyk. *Statistical analysis of network data*, volume 69 of *Springer Series in Statistics*. Springer, 2009.

[KOR15] Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):803–825, 2015.

[Li11] Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

[LW18] Olivier Ledoit and Michael Wolf. Analytical nonlinear shrinkage of large-dimensional covariance matrices. *University of Zurich, Department of Economics, Working Paper*, (264), 2018.

[Ngu11] XuanLong Nguyen. Wasserstein distances for discrete measures and convergence in nonparametric mixture models. Technical report, UNIV of Michigan MI DEPT OF STATISTICS, 2011.

[Ngu13] XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[Ngu15] XuanLong Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21(1):618–646, 2015.

[Ngu16] XuanLong Nguyen. Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure. *Bernoulli*, 22(3):1535–1571, 2016.

[Pen03] Mathew Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability*. Oxford university press, 2003.

[RBLZ08] Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[RDG08] Abel Rodriguez, David B Dunson, and Alan E Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[Res14] Sidney I Resnick. *A probability path*. Modern Birkhäuser Classics. Springer, fourth edition, 2014.

[RM11] Judith Rousseau and Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

[SM93] Elias M Stein and Timothy S Murphy. *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, volume 3 of *Monographs in Harmonic Analysis*. Princeton University Press, 1993.

[Sta65] V.A. Statulyavichus. Limit theorems for densities and asymptotic expansions for distributions of sums of independent random variables. *Theory of Probability and Its Applications*, 10(4):582–595, 1965.

[Tei67] Henry Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302, 1967.

[TJBB06] Yee W Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[Ver12] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and GittaEditors Kutyniok, editors, *Compressed Sensing: Theory and Practice*, pages 210–268. Cambridge University Press, 2012.

[Vil03] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2019.

[WVdV96] Jon Wellner and Aad Van der Vaart. *Weak convergence and empirical processes: with applications to statistics*. Springer Series in Statistics. Springer Science & Business Media, 1996.