Rates of convergence and normal approximations for estimators of local dependence random graph models

JONATHAN R. STEWART[1,a]

[1]*Department of Statistics, Florida State University, Tallahassee, FL, USA,* [a]*jrstewart@fsu.edu*

Local dependence random graph models are a class of block models for network data which allow for dependence among edges under a local dependence assumption defined around the block structure of the network. Since being introduced by Schweinberger and Handcock (2015), research in the statistical network analysis and network science literatures have demonstrated the potential and utility of this class of models. In this work, we provide the first theory for estimation and inference which ensures consistent and valid inference of parameter vectors of local dependence random graph models. This is accomplished by deriving convergence rates of estimation and inference procedures for local dependence random graph models based on a single observation of the graph, allowing both the number of model parameters and the sizes of blocks to tend to infinity. First, we derive non-asymptotic bounds on the ℓ_2-error of maximum likelihood estimators with convergence rates, outlining conditions under which these rates are minimax optimal. Second, and more importantly, we derive non-asymptotic bounds on the error of the multivariate normal approximation. These theoretical results are the first to achieve both optimal rates of convergence and non-asymptotic bounds on the error of the multivariate normal approximation for parameter vectors of local dependence random graph models.

Keywords: Local dependence random graph model; minimax bounds; multivariate normal approximation; network data; statistical network analysis

1. Introduction

Local dependence random graph models, introduced by Schweinberger and Handcock (2015), are a class of statistical models for network data built around block structure, where a population of nodes $\mathcal{N}$, which we take without loss to be $\mathcal{N} := \{1, \ldots, N\}$ ($N \geq 3$), is partitioned into $K \in \{1, 2, \ldots\}$ subsets $\mathcal{A}_1, \ldots, \mathcal{A}_K$ called blocks (also referred to as communities or subpopulations within the literature). The class owes its name to the fundamental assumption that dependence among edges is constrained to block-based subgraphs. We formally review local dependence random graph models in Section 1.1.

There are two key aspects to local dependence random graph models which help to explain the research interest received in both the statistical network analysis and network science literatures (Stewart et al., 2019, Schweinberger and Stewart, 2020, Babkin et al., 2020, Whetsell, Kroll and Dehart-Davis, 2021, Mele, 2022, Dahbura et al., 2021, Agneessens, Trincado-Munoz and Koskinen, 2024, Dahbura et al., 2023, Tolochko and Boomgaarden, 2024). First, block structure (or community structure) is a well-established structural phenomena relevant to many applications and networks encountered in our world (e.g., Holland, Laskey and Leinhardt, 1983, Newman and Girvan, 2004, Stewart et al., 2019). Second, local dependence random graph models possess desirable properties and behavior that circumvent early difficulties in constructing models of edge dependence, which include producing non-degenerate models of edge dependence (including transitivity) and consistency results for estimators (Schweinberger and Handcock, 2015, Schweinberger and Stewart, 2020).

Utilization of local dependence random graph models requires knowledge or estimates of both the block memberships of the nodes in the network, as well as the parameters of interest which determine the amount of probability mass placed on different configurations of the network. In practice, the parameter vectors of local dependence random graph models must always be estimated, whereas the block memberships of nodes can either be observed as part of the observation process (Stewart et al., 2019, Schweinberger and Stewart, 2020), or can be estimated (Babkin et al., 2020, Schweinberger, 2020). We will focus on the problem of estimating parameter vectors under the assumption that the block memberships of nodes have either been observed as part of the observation process or have been estimated.

In this work, we advance the literature on local dependence random graph models by providing the first statistical theory which elaborates conditions under which estimation and inference methodology based on a single observation of the graph can be expected to produce consistent and valid inference of parameter vectors of local dependence random graph models. The main results are non-asymptotic and cover settings where the number of model parameters and the sizes of the blocks tend to infinity. The main contributions of this work include:

1. Establishing the first non-asymptotic bounds on the ℓ_2-error of maximum likelihood estimators of parameters vectors of local dependence random graph models which hold with high probability,
2. Outlining conditions under which the rates of convergence implied by the bounds on the ℓ_2-error of maximum likelihood estimators are minimax optimal, and
3. Deriving the first non-asymptotic bound on the error of the multivariate normal approximation of a standardization of maximum likelihood estimators.

All results are stated in terms of interpretable quantities, allowing us to quantify the effect of key aspects of the statistical model and network structure upon convergence rates of the aforementioned errors. In so doing, we introduce the first principled approach to estimation and inference for local dependence random graph models by developing theoretical results which achieve both optimal rates of convergence and non-asymptotic bounds on the error of the multivariate normal approximation of maximum likelihood estimators.

1.1. Local dependence random graph models

We consider simple, undirected random graphs $\boldsymbol{X} \in \mathbb{X} \coloneqq \{0,1\}^{\binom{N}{2}}$ which are defined on the set of nodes $\mathcal{N} \coloneqq \{1,\ldots,N\}$ ($N \geq 3$). Edge variables between pairs of nodes $\{i,j\} \subset \mathcal{N}$ are given by

$$X_{i,j} \;=\; \begin{cases} 1 & \text{Nodes } i \text{ and } j \text{ are connected in the graph} \\ 0 & \text{Otherwise} \end{cases},$$

assuming throughout that $X_{i,j} = X_{j,i}$ ($\{i,j\} \subset \mathcal{N}$) and $X_{i,i} = 0$ ($i \in \mathcal{N}$).

A *local dependence random graph* (Schweinberger and Handcock, 2015) is a random graph $\boldsymbol{X}$ where the set of nodes $\mathcal{N}$ is partitioned into K *blocks* $\mathcal{A}_1,\ldots,\mathcal{A}_K$ with probability distributions $\mathbb{P}$ of the form

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) \;=\; \prod_{1 \leq k \leq l \leq K} \mathbb{P}_{k,l}(\boldsymbol{X}_{k,l} = \boldsymbol{x}_{k,l}), \qquad \boldsymbol{x} \in \mathbb{X}, \tag{1}$$

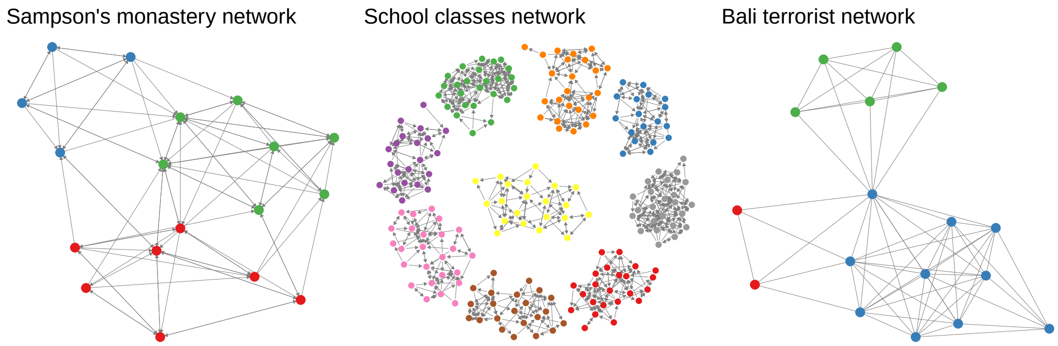

Figure 1. Three real data examples of networks for which local dependence random graph models would be applicable, including Sampson's monastery network, the school classes data set from Stewart et al. (2019), and the Bali terrorist network studied in Schweinberger and Handcock (2015). Node colors correspond to block memberships.

where the subgraphs $\boldsymbol{X}_{k,l}$ $(1 \leq k \leq l \leq K)$ are defined as follows:

$$\boldsymbol{X}_{k,l} \;\; \coloneqq \;\; \begin{cases} (X_{i,j})_{\{(i,j)\,:\, i<j,\, i \in \mathcal{A}_k, j \in \mathcal{A}_k\}} \;\in\; \mathbb{X}_{k,k} \;\coloneqq\; \{0,1\}^{\binom{|\mathcal{A}_k|}{2}} & \text{if } k = l \\ \\ (X_{i,j})_{\{(i,j)\,:\, i \in \mathcal{A}_k,\, j \in \mathcal{A}_l\}} \;\in\; \mathbb{X}_{k,l} \;\coloneqq\; \{0,1\}^{|\mathcal{A}_k|\,|\mathcal{A}_l|} & \text{if } k \neq l \end{cases}.$$

We refer to the subgraphs $\boldsymbol{X}_{k,k}$ $(1 \leq k \leq K)$ as the *within-block subgraphs* and to the subgraphs $\boldsymbol{X}_{k,l}$ $(1 \leq k < l \leq K)$ as the *between-block subgraphs*. The probability distribution $\mathbb{P}_{k,l}$ is the marginal probability distribution of the subgraph $\boldsymbol{X}_{k,l}$ $(1 \leq k \leq l \leq K)$. A *local dependence random graph model* is any probability distribution $\mathbb{P}$ for $\boldsymbol{X}$ of the form (1). Figure 1 visualizes three networks which can be studied using local dependence random graph models. While the collection of block-based subgraphs $\boldsymbol{X}_{k,l}$ $(1 \leq k \leq l \leq K)$ are independent, edges within the same block-based subgraph can be dependent. The joint distribution $\mathbb{P}$ can be specified by specifying the marginal probability distributions $\mathbb{P}_{k,l}$ for the block-based subgraphs $\boldsymbol{X}_{k,l}$ $(1 \leq k \leq l \leq K)$.

It is worth noting that the block memberships are known in both Sampson's monastery network and the school classes network visualized in Figure 1, whereas the block memberships of the Bali terrorist network were estimated as in Schweinberger and Handcock (2015). When the block memberships correspond to tangible and observable quantities (e.g., school class memberships of students), data on the block memberships can be collected as part of the observation process. When this is not possible, the block memberships must be estimated, for example by using the two-step estimation methodology of Babkin et al. (2020), which estimates both the block memberships of nodes and the parameter vectors of local dependence random graph models.

Exponential families account for the most prevalent specifications of local dependence random graph models (e.g., Schweinberger and Handcock, 2015, Stewart et al., 2019, Dahbura et al., 2021, Schubert and Brand, 2022, Tolochko and Boomgaarden, 2024), indeed having been the statistical foundations for the class in the seminal work by Schweinberger and Handcock (2015). Moreover, exponential families provide a flexible statistical platform for constructing models of edge dependence in network data applications (Lusher, Koskinen and Robins, 2012, Schweinberger et al., 2020), and have been shown to be possess desirable statistical properties in local dependence random graph models, including the consistency of maximum likelihood estimators of canonical and curved exponential families (Schweinberger and Stewart, 2020). An *exponential-family local dependence random graph model* can be specified via the marginal probability distributions of the block-based subgraphs:

$$\mathbb{P}_{k,l,\boldsymbol{\theta}_{k,l}}(\boldsymbol{X}_{k,l} = \boldsymbol{x}_{k,l}) \;=\; h_{k,l}(\boldsymbol{x}_{k,l}) \exp\left(\langle \boldsymbol{\theta}_{k,l},\, s_{k,l}(\boldsymbol{x}_{k,l})\rangle - \psi_{k,l}(\boldsymbol{\theta}_{k,l})\right), \tag{2}$$

defined for each $\boldsymbol{x}_{k,l} \in \mathbb{X}_{k,l}$, where

- $s_{k,l} : \mathbb{X}_{k,l} \mapsto \mathbb{R}^{p_{k,l}}$ is a vector of sufficient statistics;
- $\boldsymbol{\theta}_{k,l} \in \mathbb{R}^{p_{k,l}}$ is the natural parameter vector;
- $h_{k,l} : \mathbb{X}_{k,l} \mapsto [0, \infty)$ is the reference function of the exponential family; and
- $\psi_{k,l}(\boldsymbol{\theta}_{k,l}) = \log \sum_{\boldsymbol{v} \in \mathbb{X}_{k,l}} h_{k,l}(\boldsymbol{v}) \exp(\langle \boldsymbol{\theta}_{k,l},\, s_{k,l}(\boldsymbol{v})\rangle)$ is the log-normalizing constant.

It is straightforward to show that exponential family specifications of the marginal probability distributions of the within-block and between-block subgraphs will lead to a joint distribution which is also an exponential family.

A diverse range of models with the local dependence property in (1) can be constructed through different specifications of the sufficient statistics and reference functions. To allow for a general scope of well-structured models, we assume that the joint distributions of $\boldsymbol{X}$ take the form

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) \;=\; \prod_{1 \le k \le l \le K} \mathbb{P}_{k,l,\boldsymbol{\theta}_{k,l}}(\boldsymbol{X}_{k,l} = \boldsymbol{x}_{k,l}) \;=\; h(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta},\, s(\boldsymbol{x})\rangle - \psi(\boldsymbol{\theta})\right), \tag{3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_W, \boldsymbol{\theta}_B) \in \mathbb{R}^{p+q}$ and $s(\boldsymbol{x}) = (s_W(\boldsymbol{x}_W), s_B(\boldsymbol{x}_B)) \in \mathbb{R}^{p+q}$, with the definitions

$$\boldsymbol{x}_W := (\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{K,K}), \qquad \boldsymbol{x}_B := (\boldsymbol{x}_{1,2}, \ldots, \boldsymbol{x}_{1,K}, \boldsymbol{x}_{2,3}, \boldsymbol{x}_{2,4}, \ldots, \boldsymbol{x}_{K-1,K}),$$

$$h(\boldsymbol{x}) \;:=\; \prod_{1 \le k \le l \le K} h_{k,l}(\boldsymbol{x}_{k,l}), \qquad \text{and} \qquad \psi(\boldsymbol{\theta}) \;:=\; \sum_{k=1}^{K} \psi_{k,k}(\boldsymbol{\theta}_W) + \sum_{1 \le k < l \le K} \psi_{k,l}(\boldsymbol{\theta}_B).$$

Throughout, we will assume that $p = \dim(\boldsymbol{\theta}_W)$ and $q = \dim(\boldsymbol{\theta}_B)$. The exponential family is then the set of probability distributions $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^{p+q}\}$, where we note that the natural parameter space is equal to $\mathbb{R}^{p+q}$, a fact which follows trivially due to the fact that the support $\mathbb{X}$ of $\boldsymbol{X}$ is a finite set. We additionally assume throughout this work that the exponential family implied by (3) is minimal. While the product of the block-based subgraph distributions in (2) will form an exponential family, it may not be minimal, in which case we assume that the representation in (3) is the minimal representation of the exponential family obtained through reduction by sufficiency, reparameterization, and proper choice of reference measure; see Proposition 1.5 of Brown (1986). The assumption that an exponential family is minimal is not restrictive, as any non-minimal exponential family can be reduced to a minimal exponential family (Proposition 1.5, Brown, 1986).

We next provide examples of exponential-family local dependence random graph models in order to motivate the broad scope of this class of models, as well as to demonstrate how to construct local dependence random graph models. As the scope of possible models that can be constructed is large, we are unable to present a complete primer on the topic, and refer to works by Schweinberger and Handcock (2015), Stewart et al. (2019), and Schweinberger and Stewart (2020), for further information on and concrete examples of exponential-family local dependence random graph models.

1.2. Examples of exponential-family local dependence random graph models

1.2.1. Example 1: The stochastic block model

As a first example, we review the stochastic block model (Holland, Laskey and Leinhardt, 1983), which is a special case of a local dependence random graph model. The joint distribution for $\boldsymbol{X}$ is given by

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}=\boldsymbol{x}) \;&\propto\; \left[\prod_{k=1}^{K} \prod_{i<j\,:\,i,j\in\mathcal{A}_k} \exp(\theta_{k,k}\, x_{i,j})\right]\left[\prod_{1\le k<l\le K} \prod_{(i,j)\in\mathcal{A}_k\times\mathcal{A}_l} \exp(\theta_{k,l}\, x_{i,j})\right] \\
&\propto\; \exp\left(\sum_{k=1}^{K} \theta_{k,k} \sum_{i<j\,:\,i,j\in\mathcal{A}_k} x_{i,j} + \sum_{1\le k<l\le K} \theta_{k,l} \sum_{(i,j)\in\mathcal{A}_k\times\mathcal{A}_l} x_{i,j}\right),
\end{aligned}
\tag{4}
$$

where $\theta_{k,l} \in \mathbb{R}$ $(1 \le k \le l \le K)$. The second line of (4) implies the minimal exponential family, where each block-based subgraph is a collection of independent and identically distributed Bernoulli random variables whose edge probability depends on the subgraph index (k, l) and the value of $\theta_{k,l} \in \mathbb{R}$.

1.2.2. Example 2: Transitivity in local dependence random graphs

The second example we present captures stochastic tendencies towards edge transitivity in networks, by including a sufficient statistic which models the stochastic tendency for an edge in the network to belong to a triangle. For this example, we consider joint distributions $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^3\}$ for $\boldsymbol{X}$ of the form

$$
\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}=\boldsymbol{x}) \;\propto\; \exp\left(\theta_1\, s_1(\boldsymbol{x}) + \theta_2\, s_2(\boldsymbol{x}) + \theta_3\, s_3(\boldsymbol{x})\right),
$$

with natural parameters $(\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$, and where the sufficient statistics are given by

$$
\begin{aligned}
s_1(\boldsymbol{x}) \;&=\; \sum_{k=1}^{K} \sum_{i<j\,:\,i,j\in\mathcal{A}_k} x_{i,j} \\
s_2(\boldsymbol{x}) \;&=\; \sum_{k=1}^{K} \sum_{i<j\,:\,i,j\in\mathcal{A}_k} x_{i,j}\; \mathbb{1}\left(\sum_{h\in\mathcal{A}_k\setminus\{i,j\}} x_{i,h}\, x_{j,h} \;\ge\; 1\right) \\
s_3(\boldsymbol{x}) \;&=\; \sum_{1\le k<l\le K} \sum_{(i,j)\in\mathcal{A}_k\times\mathcal{A}_l} x_{i,j}.
\end{aligned}
$$

In words, $s_1(\boldsymbol{x})$ counts the number of edges in each of the within-block subgraphs $\boldsymbol{x}_{k,k}$ $(1 \le k \le K)$, whereas $s_3(\boldsymbol{x})$ counts the number of edges in each of the between-block subgraphs $\boldsymbol{x}_{k,l}$ $(1 \le k < l \le K)$. Neither of these statistics induce dependence, as when $\theta_2 = 0$, the joint distribution will factorize with respect to the edge variables in the graph, implying edges are independent.

The second sufficient statistic induces dependence among edge variables contained in the same within-block subgraph, noting that the form of the statistic in $s_2(\boldsymbol{x})$ ensures that distributions will not factorize with respect to the edge variables within the graph when $\theta_2 \neq 0$. The second statistic $s_2(\boldsymbol{x})$ counts the number of edges $x_{i,j}$ between pairs of nodes i and j belonging to a common block $\mathcal{A}_k$, which are also mutually connected to at least one other node $h \in \mathcal{A}_k$ also belonging to the same common block, i.e., it counts the number of within-block edges which form at least one triangle within the respective block-based subgraph. We call such edges *transitive edges*.

Motivation for taking this approach to constructing models of edge dependence lies in foundational properties of exponential families. The mean-value parameter map of the exponential family is given by $\boldsymbol{\mu}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\boldsymbol{\theta}}\,(s_1(\boldsymbol{X}), s_2(\boldsymbol{X}), s_3(\boldsymbol{X}))$ (p. 73–74, Brown, 1986), mapping the natural parameter space $\mathbb{R}^3$ to the interior of the convex hull of the image of $\mathbb{X}$ under the vector of sufficient statistics $s : \mathbb{X} \mapsto \mathbb{R}^3$:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) \quad \in \quad \mathbb{M} \quad \coloneqq \quad \operatorname{int}\left(\operatorname{ConHull}\left(\left\{s(\boldsymbol{x}) \in \mathbb{R}^3 \,:\, \boldsymbol{x} \in \mathbb{X}\right\}\right)\right),$$

where $\operatorname{ConHull}(\mathbb{S})$ represents the convex hull of the set $\mathbb{S}$. Moreover, for a minimal exponential family (of which this example is), the map $\boldsymbol{\mu} : \mathbb{R}^3 \mapsto \mathbb{M}$ defines a homeomorphism between $\mathbb{R}^3$ and $\mathbb{M}$ (Theorem 3.6, Brown, 1986). This last point emphasizes a key modeling aspect of exponential-family local dependence random graph models, as for any point $\boldsymbol{u} \in \mathbb{M}$ parameterizing the expected values (mean values) of the sufficient statistics $(s_1(\boldsymbol{X}), s_2(\boldsymbol{X}), s_3(\boldsymbol{X}))$, we are guaranteed to be able to find a natural parameter vector $\boldsymbol{\theta} \in \mathbb{R}^3$ for which $\mathbb{E}_{\boldsymbol{\theta}}\,(s_1(\boldsymbol{X}), s_2(\boldsymbol{X}), s_3(\boldsymbol{X})) = \boldsymbol{u}$, allowing specified models to flexibly capture average tendencies of networks, including density, transitivity, and much more.

1.2.3. Example 3: Incorporating node and block heterogeneity into models

The third example shows how we are able to incorporate heterogeneous parameterizations for blocks, as well as for the stochastic propensities of different nodes to form edges, demonstrating ways in which the dimension of parameter vectors can grow in applications. For ease of presentation, we will build on Example 2 by extending the sufficient statistics which were specified in that example.

First, we will demonstrate how heterogeneity in node degrees can be incorporated into models. Suppose that nodes are divided into M non-overlapping groups or categories $\{1, \ldots, M\}$ which we represent as sets $\mathscr{G}_1, \ldots, \mathscr{G}_M$. Note that these are distinct from the blocks $\mathcal{A}_1, \ldots, \mathcal{A}_K$. In applications, these groups might be ranks in a department, gender, race, or any other categorical covariate which can be observed and treated as fixed. As such, each block may be comprised of different amount of nodes from each of the groups $\mathscr{G}_1, \ldots, \mathscr{G}_M$, an example of which is the school classes data set studied in Stewart et al. (2019), where each school class was comprised of different numbers of male and female students.

We replace $s_1(\boldsymbol{x})$ in Example 2 by multiple sufficient statistics:

$$s_m(\boldsymbol{x}) \quad = \quad \sum_{k=1}^{K} \sum_{i \in \mathscr{G}_m \cap \mathcal{A}_k} \sum_{j \in \mathcal{A}_k \setminus \{i\}} x_{i,j}, \qquad m \in \{1, \ldots, M\},$$

with natural parameters $(\theta_1, \ldots, \theta_M) \in \mathbb{R}^M$. A version of this statistic is implemented in the `R` package `ergm` under the name `nodefactor` (Krivitsky et al., 2023). In words, the model includes a sufficient statistic that, based on the value of the corresponding natural parameter, adjusts the baseline propensity for within-block edge formation involving nodes in that group. With no other sufficient statistics in the model, the log-odds of an edge would be given by

$$\log \frac{\mathbb{P}_{\boldsymbol{\theta}}(X_{i,j} = 1)}{\mathbb{P}_{\boldsymbol{\theta}}(X_{i,j} = 0)} \quad = \quad \theta_m + \theta_n, \qquad i \in \mathscr{G}_m \cap \mathcal{A}_k,\; j \in \mathscr{G}_n \cap \mathcal{A}_k, \qquad k \in \{1, \ldots, K\}.$$

This is reminiscent of the $p1$ model (Holland and Leinhardt, 1981) and the β-model (Chatterjee, Diaconis and Sly, 2011), in which each node is given its own distinct class.

We now show how heterogeneity can arise in the block-based subgraphs, by allowing different parameterizations for different blocks. The statistic $s_2(\boldsymbol{x})$ in Example 2 counts the number of transitive edges in each within-block subgraph $\boldsymbol{X}_{k,k}$ $(1 \leq k \leq K)$, using the same parameter for each within-block subgraph. It may be that different blocks display different tendencies towards transitivity. To make this concrete, suppose that the individual blocks $\{1, \ldots, K\}$ are partitioned into L groups or categories $\mathscr{H}_1, \ldots, \mathscr{H}_L$. We then replace the sufficient statistic $s_2(\boldsymbol{x})$ in Example 2 by multiple statistics:

$$s_{M+l}(\boldsymbol{x}) \;\; = \;\; \sum_{k \in \mathscr{H}_l} \; \sum_{i<j\,:\, i,j \in \mathcal{A}_k} x_{i,j} \; \mathbb{1}\left(\sum_{h \in \mathcal{A}_k \setminus \{i,j\}} x_{i,h}\, x_{j,h} \;\geq\; 1 \right), \qquad l \in \{1, \ldots, L\},$$

with natural parameters $(\theta_{M+1}, \ldots, \theta_{M+L}) \in \mathbb{R}^L$. The complete model is given by

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) \;\; \propto \;\; \exp\left(\sum_{t=1}^{M+L+1} \theta_t \; s_t(\boldsymbol{x}) \right),$$

with natural parameter space $\mathbb{R}^{M+L+1}$, where the last sufficient statistic is equal to

$$s_{M+L+1}(\boldsymbol{x}) \;\; = \;\; \sum_{1 \leq k < l \leq K} \; \sum_{(i,j) \in \mathcal{A}_k \times \mathcal{A}_l} x_{i,j}.$$

Example 3 helps to demonstrate how the number of model parameters can grow quickly in applications when significant generality, heterogeneity, or adaptability is needed to capture important aspects of the application. A version of Example 3 will be used in the simulation studies conducted in Section 3.

2. Theoretical guarantees

Our main theoretical results are presented in this section. We first review exponential family theory for local dependence random graph models in Section 2.1. Our consistency theory is then presented in Section 2.2. Section 2.2.1 derives rates of convergence in the ℓ_2-norm of maximum likelihood estimators, whereas Section 2.2.2 presents bounds on the minimax risk in the ℓ_2-norm which help to establish the minimax optimality (under mild conditions) of the upper bounds presented in Section 2.2.1. Lastly, but importantly, rates of convergence of the error of the multivariate normal approximation are obtained in Section 2.3, providing both non-asymptotic and asymptotic theory for multivariate normal approximations of maximum likelihood estimators of local dependence random graph models.

Due to space restrictions, all proofs are presented in the supplement (Stewart, 2024).

2.1. Preliminaries for exponential families

The log-likelihood of an exponential-family local dependence random graph model is

$$\ell(\boldsymbol{\theta}, \boldsymbol{x}) \;\; := \;\; \log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) \;\; = \;\; \sum_{k=1}^{K} \ell_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) + \sum_{1 \leq k < l \leq K} \ell_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}),$$

where

$$\begin{aligned} \ell_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) \;\; &:= \;\; \langle \boldsymbol{\theta}_W,\, s_{k,k}(\boldsymbol{x}_{k,k}) \rangle - \psi_{k,k}(\boldsymbol{\theta}_W) + \log h_{k,k}(\boldsymbol{x}_{k,k}) \\ \ell_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}) \;\; &:= \;\; \langle \boldsymbol{\theta}_B,\, s_{k,l}(\boldsymbol{x}_{k,l}) \rangle - \psi_{k,l}(\boldsymbol{\theta}_B) + \log h_{k,l}(\boldsymbol{x}_{k,l}). \end{aligned}$$

The gradient $\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{x}) = (\nabla_{\boldsymbol{\theta}_W}\, \ell(\boldsymbol{\theta}, \boldsymbol{x}), \nabla_{\boldsymbol{\theta}_B}\, \ell(\boldsymbol{\theta}, \boldsymbol{x}))$ is given by

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_W}\, \ell(\boldsymbol{\theta}, \boldsymbol{x}) &= \sum_{k=1}^{K} \left[s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}_{k,k,\boldsymbol{\theta}_W}\, s_{k,k}(\boldsymbol{X}_{k,k})\right] \\ \nabla_{\boldsymbol{\theta}_B}\, \ell(\boldsymbol{\theta}, \boldsymbol{x}) &= \sum_{1 \le k < l \le K} \left[s_{k,l}(\boldsymbol{x}_{k,l}) - \mathbb{E}_{k,l,\boldsymbol{\theta}_B}\, s_{k,l}(\boldsymbol{X}_{k,l})\right],\end{aligned}$$

where $\mathbb{E}_{k,k,\boldsymbol{\theta}_W}$ and $\mathbb{E}_{k,l,\boldsymbol{\theta}_B}$ are the expectation operators with respect to the marginal probability distributions $\mathbb{P}_{k,k,\boldsymbol{\theta}_W}$ of $\boldsymbol{X}_{k,k}$ and $\mathbb{P}_{k,l,\boldsymbol{\theta}_B}$ of $\boldsymbol{X}_{k,l}$, respectively (Lemma 6.1, Stewart, 2024). We denote the set of maximum likelihood estimators for a given observation $\boldsymbol{x} \in \mathbb{X}$ by

$$\widehat{\boldsymbol{\Theta}} \;\equiv\; \widehat{\boldsymbol{\Theta}}(\boldsymbol{x}) \;:=\; \left\{\boldsymbol{\theta}' \in \mathbb{R}^{p+q} \,:\, \ell(\boldsymbol{\theta}', \boldsymbol{x}) = \sup_{\boldsymbol{\theta} \in \mathbb{R}^{p+q}} \ell(\boldsymbol{\theta}, \boldsymbol{x})\right\}.$$

For minimal and regular exponential families, the maximum likelihood estimator exists uniquely when it is exists, i.e., $|\widehat{\boldsymbol{\Theta}}| \in \{0, 1\}$ (Proposition 3.13, Sundberg, 2019). Regarding existence, the maximum likelihood estimator of natural parameter vectors of minimal exponential families exists when the sufficient statistic vector falls within the interior of the mean-value parameter space (Theorem 5.5., p. 148, Brown, 1986), in which case there exists a parameter vector $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^{p+q}$ for which $\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}) = s(\boldsymbol{x})$ for a given observation $\boldsymbol{x} \in \mathbb{X}$ of the random graph $\boldsymbol{X}$, defining $\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{X})$ to be the mean-value parameter map (p. 73–74, Brown, 1986).

In practice, computing maximum likelihood estimators is not straightforward, as the log-normalizing constants are generally computationally intractable unless the marginal probability distributions of the block-based subgraphs $\boldsymbol{X}_{k,l}$ $(1 \le k \le l \le K)$ are assumed to factorize further to reduce the computational burden of computing the normalizing constants, because $\psi_{k,k}(\boldsymbol{\theta}_W)$ involves the summation of $\binom{|\mathcal{A}_k|}{2}$ terms and $\psi_{k,l}(\boldsymbol{\theta}_B)$ involves the summation of $|\mathcal{A}_k|\,|\mathcal{A}_l|$ terms. It becomes infeasible to compute these summations in practice even for modest block sizes. The prevailing method for estimating exponential families of random graph models is Monte-Carlo maximum likelihood estimation (MCMLE) (Hunter and Handcock, 2006). The algorithm outlined in Hunter and Handcock (2006) applies directly to exponential-family local dependence random graph models (Stewart and Schweinberger, 2019, Stewart et al., 2019, Schweinberger and Stewart, 2020), and is used in the simulation studies conducted in Section 3 through the implementation in the `R` package `mlergm` (Stewart and Schweinberger, 2019).

We summarize the key aspects of MCMLE with exponential families of random graph models outlined in Hunter and Handcock (2006). The essential idea of MCMLE is to approximate intractable likelihood functions with stochastic approximations utilizing Markov Chain Monte Carlo (MCMC) methods. The crux of the methodology rests on a simple approximation of normalizing constants via importance sampling. To introduce the idea, let $\boldsymbol{\theta}_0 \in \mathbb{R}^{p+q}$ be a fixed parameter vector in the natural parameter space of an exponential-family local dependence random graph model. We can equivalently find maximum likelihood estimators $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^{\star}$ by

$$\widehat{\boldsymbol{\theta}} \;=\; \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \mathbb{R}^{p+q}} \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) - \ell(\boldsymbol{\theta}_0, \boldsymbol{x})\right] \;=\; \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \mathbb{R}^{p+q}} \left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0,\, s(\boldsymbol{x})\rangle - \log\left(\exp\left(\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)\right)\right)\right].$$

In order to solve the above optimization problem, we need to be able to approximate the gradient corresponding to the above objective function, which is given by

$$\nabla_{\boldsymbol{\theta}}\, \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) - \ell(\boldsymbol{\theta}_0, \boldsymbol{x})\right] \;=\; s(\boldsymbol{x}) - \nabla_{\boldsymbol{\theta}} \log\left(\exp\left(\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)\right)\right). \tag{5}$$

The intractability of the normalizing constants in (5) makes direct computation infeasible, as discussed.

We approximate the term $\exp(\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0))$ via a change of measure argument:

$$\begin{aligned}
\exp(\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)) &= \exp(-\psi(\boldsymbol{\theta}_0)) \sum_{\boldsymbol{x} \in \mathbb{X}} h(\boldsymbol{x}) \exp(\langle \boldsymbol{\theta}, s(\boldsymbol{x}) \rangle) \\
&= \exp(-\psi(\boldsymbol{\theta}_0)) \sum_{\boldsymbol{x} \in \mathbb{X}} h(\boldsymbol{x}) \exp(\langle \boldsymbol{\theta}, s(\boldsymbol{x}) \rangle) \frac{\exp(\langle \boldsymbol{\theta}_0, s(\boldsymbol{x}) \rangle)}{\exp(\langle \boldsymbol{\theta}_0, s(\boldsymbol{x}) \rangle)} \\
&= \mathbb{E}_{\boldsymbol{\theta}_0} \exp(\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, s(\boldsymbol{X}) \rangle),
\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{\theta}_0}$ is the expectation operator corresponding to $\mathbb{P}_{\boldsymbol{\theta}_0}$. As a result, if we can approximate the expectation $\mathbb{E}_{\boldsymbol{\theta}_0} \exp(\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, s(\boldsymbol{X}) \rangle)$ via Monte Carlo methods, then we can approximate the ratio of normalizing constants $\exp(\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0))$. A key advantage of this approach lies in the fact that the expectation is taken with respect to a fixed distribution $\mathbb{P}_{\boldsymbol{\theta}_0}$. In general, we will not be able to sample directly from the distributions and will need to rely on MCMC sampling methods (see, e.g., Snijders, 2002, Krivitsky et al., 2023). Let $\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_n$ be an MCMC sample from $\mathbb{P}_{\boldsymbol{\theta}_0}$. Then, returning to (5), we have the following approximation:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) - \ell(\boldsymbol{\theta}_0, \boldsymbol{x})\right] &\approx s(\boldsymbol{x}) - \nabla_{\boldsymbol{\theta}} \log \left(\frac{1}{n} \sum_{i=1}^{n} \exp(\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, s(\widetilde{\boldsymbol{x}}_i) \rangle) \right) \\
&= s(\boldsymbol{x}) - \sum_{i=1}^{n} \left(\frac{\exp(\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, s(\widetilde{\boldsymbol{x}}_i) \rangle)}{\sum_{j=1}^{n} \exp\left(\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, s(\widetilde{\boldsymbol{x}}_j) \rangle\right)} \right) s(\widetilde{\boldsymbol{x}}_i).
\end{aligned}$$

Using this approximation, root finding algorithms—such as stochastic gradient descent or Fisher scoring algorithms—can be utilized to find the MCMLE approximation to the MLE; Hunter and Handcock (2006) outlines a stochastic Fisher scoring algorithm. The convergence of the MCMLE to the MLE depends on the convergence of the exact log-likelihood to the stochastic approximation (see, e.g., discussions in Geyer and Thompson, 1992), which will depend upon properties of the Markov chain utilized to generate sample networks. In the usual implementations (e.g., Krivitsky et al., 2023), these chains will be geometrically mixing toward the target sampling distribution and will provide good approximations provided sufficient computational resources have been expended. As a final point on the computational complexity, different model specifications, implementations of MCMC methodology, and block structures will have different mixing times and thus will require differing amounts of computational resources. With regards to scalability, access to parallel computing presents a significant opportunity to improve computation times by exploiting the independence of the block-based subgraphs to parallelize simulation; see discussions in Babkin et al. (2020), which analyzed networks with over 10,000 nodes utilizing parallel computing and an implementation of the stochastic Fisher scoring algorithm of Hunter and Handcock (2006) in the `R` package `mlergm` (Stewart and Schweinberger, 2019).

2.2. Convergence rates of maximum likelihood estimators

We derive non-asymptotic bounds on the ℓ_2-error of maximum likelihood estimators which hold with high probability. Our results extend those of Schweinberger and Stewart (2020), who derived consistency results for maximum likelihood estimators of canonical and curved exponential-family local dependence random graph models, but did not report rates of convergence. Additionally, Schweinberger and Stewart (2020) focused on estimation of only the within-block parameter vectors $\boldsymbol{\theta}_W$. In contrast, we establish consistency theory with rates of convergence for entire parameter vectors $(\boldsymbol{\theta}_W, \boldsymbol{\theta}_B)$ of exponential-family local dependence random graph models, covering settings where the number of model parameters and sizes of blocks may tend to infinity, at appropriate rates. The consistency theory in this work is related to—but distinct from—the results in Stewart and Schweinberger (2020), who prove a general theorem for establishing consistency and rates of convergence of maximum likelihood and pseudolikelihood-based estimators of random graph models with dependent edges with respect to the ℓ_∞-norm under a more general weak dependence assumption. First, we focus specifically on local dependence random graph models and quantify rates of convergence in the ℓ_2-norm for this class of models and in terms of interpretable quantities related to local dependence random graphs, namely properties of the block structure, graph, and model. Second, our method of proof is fundamentally different from that of both Schweinberger and Stewart (2020) and Stewart and Schweinberger (2020), and the consistency theory in this work cannot be proved as a corollary to an existing result.

We outline some notational definitions and regularity assumptions for our theorems to follow, subsequently discussing each in turn. Let $\mathcal{B}_2(\boldsymbol{v}, r) := \{\boldsymbol{v}' \in \mathbb{R}^{\dim(\boldsymbol{v})} : \|\boldsymbol{v}' - \boldsymbol{v}\|_2 < r\}$ be the open ℓ_2-ball with center $\boldsymbol{v}$ and radius $r > 0$ and denote by $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ the smallest and largest eigenvalues, respectively, of the matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$. We write $a_N = O(b_N)$ when there exists a constant $C > 0$ and integer $N_0 \geq 1$ such that $a_N \leq C\, b_N$ for all $N \geq N_0$, and write $a_N = o(b_N)$ when there exists, for all $\delta > 0$, an integer $N_0(\delta) \geq 1$ such that $a_N \leq \delta\, b_N$ for all $N \geq N_0(\delta)$.

Assumption 1. Assume there exist $C_W > 0$ and $C_B > 0$, independent of N, p, and q, such that

$$\begin{aligned} \sup_{\boldsymbol{x}_{k,k} \in \mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k})\|_\infty &\leq C_W \binom{|\mathcal{A}_k|}{2}, \qquad k \in \{1, \ldots, K\}, \\ \sup_{\boldsymbol{x}_{k,l} \in \mathbb{X}_{k,l}} \|s_{k,l}(\boldsymbol{x}_{k,l})\|_\infty &\leq C_B\, |\mathcal{A}_k|\, |\mathcal{A}_l|, \qquad \{k, l\} \subseteq \{1, \ldots, K\}. \end{aligned}$$

Assumption 2. Assume there exists $\epsilon > 0$, independent of N, p, and q, such that

$$\begin{aligned} \widetilde{\lambda}_{\min,W}^{\epsilon} &:= \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \frac{\lambda_{\min}\left(-\mathbb{E}\, \nabla_{\boldsymbol{\theta}_W}^2\, \ell(\boldsymbol{\theta}, \boldsymbol{X})\right)}{K} > 0 \\ \widetilde{\lambda}_{\min,B}^{\epsilon} &:= \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \frac{\lambda_{\min}\left(-\mathbb{E}\, \nabla_{\boldsymbol{\theta}_B}^2\, \ell(\boldsymbol{\theta}, \boldsymbol{X})\right)}{\binom{K}{2}} > 0. \end{aligned}$$

Assumption 3. Define $A_{\text{avg}} := K^{-1} \sum_{k=1}^{K} |\mathcal{A}_k|$ to be the average block size and

$$\widetilde{\lambda}_{\max,W}^{\star} := \frac{\lambda_{\max}\left(-\mathbb{E}\, \nabla_{\boldsymbol{\theta}_W}^2\, \ell(\boldsymbol{\theta}^\star, \boldsymbol{X})\right)}{K} \quad \text{and} \quad \widetilde{\lambda}_{\max,B}^{\star} := \frac{\lambda_{\max}\left(-\mathbb{E}\, \nabla_{\boldsymbol{\theta}_B}^2\, \ell(\boldsymbol{\theta}^\star, \boldsymbol{X})\right)}{\binom{K}{2}},$$

and assume that

$$\sqrt{A_{\text{avg}}}\, \frac{\sqrt{\widetilde{\lambda}_{\max,W}^{\star}}}{\widetilde{\lambda}_{\min,W}^{\epsilon}} = o\left(\sqrt{\frac{N}{p}}\right) \quad \text{and} \quad A_{\text{avg}}\, \frac{\sqrt{\widetilde{\lambda}_{\max,B}^{\star}}}{\widetilde{\lambda}_{\min,B}^{\epsilon}} = o\left(\sqrt{\frac{N^2}{q}}\right).$$

Assumption 4. Assume the largest block size $A_{\max} := \max\{|\mathcal{A}_1|, \ldots, |\mathcal{A}_K|\}$ satisfies

$$A_{\max} \leq \min\left\{\left(\frac{N\, \widetilde{\lambda}_{\max,W}^{\star}}{A_{\text{avg}}\, p^2}\right)^{1/4}, \left(\frac{N^2\, \widetilde{\lambda}_{\max,B}^{\star}}{4\, A_{\text{avg}}^2\, q^2}\right)^{1/4}\right\}.$$

Remark 1 (Discussion of Assumption 1). We place a restriction on the scaling of the block-based sufficient statistic vectors with respect to the sizes of the blocks. The need for this arises out of a need to derive concentration inequalities for gradients of the log-likelihood, as well as a need to control third-order derivatives of the log-likelihood function in our method of proof for deriving bounds on the error of the multivariate normal approximation. The assumption is natural, as it essentially requires that the values of the sufficient statistics possess an upper-bound which is proportional to the number of edge variables in each of the respective block-based subgraphs. An example of interest is the transitive edge count statistic of a within-block subgraph $\boldsymbol{X}_{k,k}$, discussed also in in Section 1.2, given by

$$\sum_{i<j \,:\, i,j \in \mathcal{A}_k} x_{i,j} \; \mathbb{1}\left(\sum_{h \in \mathcal{A}_k \setminus \{i,j\}} x_{i,h}\, x_{j,h} \;\geq\; 1 \right) \;\leq\; \sum_{i<j \,:\, i,j \in \mathcal{A}_k} x_{i,j} \;\leq\; \binom{|\mathcal{A}_k|}{2},$$

which can be viewed as a special case of the geometrically-weighted edgewise shared partner statistic (Hunter and Handcock, 2006, Stewart et al., 2019). To further contextualize this assumption, it is helpful to note that Assumption 1 is related to the issue of instability of exponential-families of random graph models (Schweinberger, 2011). Maximal changes in the sufficient statistic vectors $s_{k,k}(\boldsymbol{x})$ $(1 \leq k \leq K)$ and $s_{k,l}(\boldsymbol{x})$ $(1 \leq k < l \leq K)$ due to changing the value of a single edge in $\boldsymbol{x}$ are defining characteristics of instability in exponential-family random graph models, in the sense of Schweinberger (2011). Assumption 1 implies limitations on the sensitivity of the sufficient statistic vectors to changes in the edges in the graph. Understanding this connection helps to explain why local dependence random graph models achieve statistical behavior and properties not achieved in early—but flawed—statistical models of edge dependence in network data (Häggström and Jonasson, 1999, Jonasson, 1999, Schweinberger, 2011, Chatterjee and Diaconis, 2013). Lastly, it is worth noting that Assumption 1 could be relaxed further, allowing for a larger upper bound. The result of this, however, would be looser upper-bounds on the ℓ_2-error and slower rates of convergence.

Remark 2 (Discussion of Assumption 2). Assumption 2 places a restriction on the scaling of the smallest eigenvalue of the joint Fisher information matrix by placing an assumption on the scaling of the smallest eigenvalue of the Fisher information matrices $-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_W}\, \ell(\boldsymbol{\theta}, \boldsymbol{X})$ and $-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_B}\, \ell(\boldsymbol{\theta}, \boldsymbol{X})$ corresponding to the within-block and between-block probability distributions, respectively, in a neighborhood $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ of the data-generating parameter vector $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}^\star_W, \boldsymbol{\theta}^\star_B)$. The local dependence assumption and the assumption that the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_W, \boldsymbol{\theta}_B) \in \mathbb{R}^{p+q}$ partitions the within-block and between-block parameters implies that the joint Fisher information matrix $-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{X})$ has the form

$$-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{X}) \;=\; \begin{pmatrix} -\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_W}\, \ell(\boldsymbol{\theta}, \boldsymbol{X}) & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & -\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_B}\, \ell(\boldsymbol{\theta}, \boldsymbol{X}) \end{pmatrix},$$

where $\mathbf{0}_{m,n}$ is the $(m \times n)$-dimensional matrix of all zeros. Assumption 2 essentially assumes that the Fisher information matrices are invertible in a neighborhood of the data-generating parameter vector. Minimum eigenvalue restrictions of Fisher information matrices are standard in settings where the number of model parameters may tend to infinity (e.g., Portnoy, 1988, Ravikumar, Wainwright and Lafferty, 2010, Janková and van de Geer, 2018). Notably, our assumption represents a restriction on what amounts to an average minimum eigenvalue (averaged over the block-based quantities in both the within-block and between-block cases). To understand why we have adopted this definition in our assumptions (relevant also to Assumption 3), instead of placing a restriction on the minimum eigenvalue of the Fisher information matrices corresponding to each block-based subgraph, observe through Weyl's inequality, the bound

$$\begin{aligned}\lambda_{\min}\left(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell(\boldsymbol{\theta},\boldsymbol{X})\right) &\geq \sum_{k=1}^{K}\lambda_{\min}\left(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell_{k,k}(\boldsymbol{\theta}_W,\boldsymbol{X}_{k,k})\right)\\ &\geq K\left(\min_{k\in\{1,\ldots,K\}}\lambda_{\min}\left(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell_{k,k}(\boldsymbol{\theta}_W,\boldsymbol{X}_{k,k})\right)\right).\end{aligned}$$

If certain subgraphs do not contain any information about certain subsets of parameters, possibly due to heterogeneous parameterizations that allow different blocks to have different parameters, then it may be the case that $\lambda_{\min}(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell_{k,k}(\boldsymbol{\theta}_W,\boldsymbol{X}_{k,k})) = 0$ for some $k \in \{1,\ldots,K\}$ due to singularity. As a result and in order to cover more general settings and heterogeneous parameterizations, we place our minimum eigenvalue restriction on the scaling of the averaged smallest eigenvalue of joint Fisher information matrices.

Remark 3 (Discussion of Assumption 3). Assumption 3 places a regularity assumption on three key quantities, the average block size $A_{\text{avg}} := K^{-1}\sum_{k=1}^{K}|\mathcal{A}_k|$, the average minimum eigenvalues of Fisher information matrices $\widetilde{\lambda}^{\epsilon}_{\min,W}$ and $\widetilde{\lambda}^{\epsilon}_{\min,B}$ in a neighborhood of the data-generating parameter vector $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}^\star_W,\boldsymbol{\theta}^\star_B)$ (defined in Assumption 2), and the average maximum eigenvalues of Fisher information matrices $\widetilde{\lambda}^{\star}_{\max,W}$ and $\widetilde{\lambda}^{\star}_{\max,B}$ at the data-generating parameter vector. As will be seen in Theorem 2.1, Assumption 3 essentially outlines a scaling requirement of these three quantities (in their respective cases) which ensures consistent estimation under Theorem 2.1, in the sense that the upper bounds on the ℓ_2-error in Theorem 2.1 will tend to zero as the size of the network N tends to infinity. As such, Assumption 3 can be viewed as a minimal information criterion which requires that we obtain sufficient information about the parameter vector $(\boldsymbol{\theta}^\star_W,\boldsymbol{\theta}^\star_B)$ from an observation of the random graph $\boldsymbol{X}$.

Remark 4 (Discussion of Assumption 4). In our method of deriving concentration inequalities, we bound factors involving the influence of edge variables in the random graph by the size of the largest block size, noting that dependence is restricted to block-based subgraphs whose size is dominated by functions of the largest block size. Similar approaches have been taken in Schweinberger and Stewart (2020). Notably, Assumption 4 does not assume that the sizes of blocks are fixed and allows these quantities to grow without bound. However, this assumption places a restriction on how large blocks can be in order to ensure that the derived concentration inequalities are sufficiently sharp to facilitate the development of the statistical theory of this work.

2.2.1. Upper bounds on the ℓ_2-error of maximum likelihood estimators

The first theoretical result we present establishes upper-bounds on the ℓ_2-error of maximum likelihood estimators for exponential-family local dependence random graph models which hold with high probability, presented in Theorem 2.1. This paves the way for establishing bounds on rates of convergence of maximum likelihood estimators with respect to the ℓ_2-norm. We will address the question of optimal rates of convergence in Section 2.2.2, where we outline a set of sufficient conditions for which we prove the upper bounds in Theorem 2.1 are minimax optimal, in the sense that the upper bounds derived in Theorem 2.1 match (up to an unknown constant) the minimax rate of convergence.

Theorem 2.1. *Consider a minimal exponential-family local dependence random graph model satisfying Assumptions 1, 2, 3, and 4 and assume that $p = \dim(\boldsymbol{\theta}_W^\star) \geq \log N$ and $q = \dim(\boldsymbol{\theta}_B^\star) \geq \log N$. Then there exist constants $C > 0$ and $N_0 \geq 3$, independent of N, p, and q, such that, with probability at least $1 - N^{-2}$, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_W, \widehat{\boldsymbol{\theta}}_B) \in \mathbb{R}^{p+q}$ exists, is unique, and satisfies*

$$\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_2 \;\leq\; C\,\sqrt{A_{\text{avg}}}\,\frac{\sqrt{\widetilde{\lambda}^\star_{\max,W}}}{\widetilde{\lambda}^\epsilon_{\min,W}}\,\sqrt{\frac{p}{N}}$$

$$\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_2 \;\leq\; C\,A_{\text{avg}}\,\frac{\sqrt{\widetilde{\lambda}^\star_{\max,B}}}{\widetilde{\lambda}^\epsilon_{\min,B}}\,\sqrt{\frac{q}{N^2}},$$

for all integers $N \geq N_0$.

Theorem 2.1 provides the foundation for establishing convergence rates in the ℓ_2-norm of maximum likelihood estimators of exponential-family local dependence random graph models. The assumption that the exponential family is minimal ensures uniqueness of the maximum likelihood estimator when it exists (Proposition 3.13, Sundberg, 2019). Rates of convergence will depend on

- the dimensions of the parameters vectors $p = \dim(\boldsymbol{\theta}_W^\star)$ and $q = \dim(\boldsymbol{\theta}_B^\star)$;
- the ratios $\sqrt{\widetilde{\lambda}^\star_{\max,W}} \,/\, \widetilde{\lambda}^\epsilon_{\min,W}$ and $\sqrt{\widetilde{\lambda}^\star_{\max,B}} \,/\, \widetilde{\lambda}^\epsilon_{\min,B}$; and
- the average block size A_{avg},

with rates of convergence depending on the scaling of these quantities with respect to N.

We permit both $\widetilde{\lambda}^\star_{\max,W}$ and $\widetilde{\lambda}^\star_{\max,B}$ to scale faster than $\widetilde{\lambda}^\epsilon_{\min,W}$ and $\widetilde{\lambda}^\epsilon_{\min,B}$, respectively, provided consistency is still established (i.e., provided Assumption 3 is met). Within the context of exponential families of growing dimension in classical settings of a random sample of independent and identically distributed random vectors, Portnoy (1988) and Ghosal (2000) obtain similar convergence rates, in their respective settings. Notably, Theorem 2.1 of Portnoy (1988) arrives at a similar scaling requirement for the minimum and maximum eigenvalues of Fisher information matrices. A key difference is that both works place third order assumptions on the models (see the assumptions of Theorem 2.1 of Portnoy (1988), and Theorem 2.1 of Ghosal (2000)). We avoid the need for such assumptions through the method of proof of Theorem 2.1, but require a smoothness condition on minimum eigenvalues of Fisher information matrices, as Assumptions 2 and 3 restrict the scaling of maximum eigenvalues of the Fisher information matrix at the data-generating parameter vector $\boldsymbol{\theta}^\star$ relative to minimum eigenvalues of the same within a neighborhood $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ of $\boldsymbol{\theta}^\star$. If we assume additional regularity in the spectrum of the Fisher information matrices by assuming that

$$\widetilde{\lambda}^\star_{\min,W} \;:=\; \frac{\lambda_{\min}\left(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell(\boldsymbol{\theta}^\star, \boldsymbol{X})\right)}{K} \;=\; O\left(\widetilde{\lambda}^\epsilon_{\min,W}\right)$$

$$\widetilde{\lambda}^\star_{\min,B} \;:=\; \frac{\lambda_{\min}\left(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_B}\,\ell(\boldsymbol{\theta}^\star, \boldsymbol{X})\right)}{\binom{K}{2}} \;=\; O\left(\widetilde{\lambda}^\epsilon_{\min,B}\right),$$

then we could prove a corollary to Theorem 2.1 which establishes the upper bounds

$$\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}^\star_W\|_2 \;\leq\; C\,\sqrt{A_{\mathrm{avg}}}\,\frac{\sqrt{\widetilde{\lambda}^\star_{\max,W}}}{\widetilde{\lambda}^\star_{\min,W}}\,\sqrt{\frac{p}{N}}$$

$$\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}^\star_B\|_2 \;\leq\; C\,A_{\mathrm{avg}}\,\frac{\sqrt{\widetilde{\lambda}^\star_{\max,B}}}{\widetilde{\lambda}^\star_{\min,B}}\,\sqrt{\frac{q}{N^2}},$$

which are more analogous to the results of Portnoy (1988).

As a final point, Theorem 2.1 assumes that the block memberships are known, i.e., the blocks $\mathcal{A}_1, \ldots, \mathcal{A}_K$ are observed or estimated without error. In many cases, the block memberships can be observed through the observation process (e.g., Stewart et al., 2019, Schweinberger and Stewart, 2020). However, in certain settings this may not be possible and the block memberships must be estimated (e.g., Babkin et al., 2020, Schweinberger, 2020). In both cases, the results of Theorem 2.1 can be regarded as the estimation error of an oracle estimate with perfect knowledge or estimation of the block structure of the network. The impact of imperfect block membership knowledge on theoretical guarantees (whether through a noisy observation or error in the estimation of block memberships of nodes) is an open question for future research.

2.2.2. Minimax risk in the ℓ_2-norm and optimal rates of convergence

We next turn to the question of whether the upper bounds on the ℓ_2-error established in Theorem 2.1 are optimal, in the sense that they match (up to an unknown constant) the rates of convergence of the minimax risk in the ℓ_2-norm.

We define the minimax risk with respect to the ℓ_2-norm to be

$$\begin{aligned}\mathscr{R}_{W,N} &:= \inf_{\widehat{\boldsymbol{\theta}}_W}\ \sup_{\boldsymbol{\theta}\in\mathbb{R}^{p+q}}\ \mathbb{E}_{\boldsymbol{\theta}}\,\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \\ \mathscr{R}_{B,N} &:= \inf_{\widehat{\boldsymbol{\theta}}_B}\ \sup_{\boldsymbol{\theta}\in\mathbb{R}^{p+q}}\ \mathbb{E}_{\boldsymbol{\theta}}\,\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B\|_2.\end{aligned} \tag{6}$$

The method by which we establish lower bounds to the minimax risk in the ℓ_2-norm requires placing an assumptions on the average value of the largest eigenvalues of Fisher information matrices, similar to the roles of $\widetilde{\lambda}^\star_{\max,W}$ and $\widetilde{\lambda}^\star_{\max,B}$ in Theorem 2.1, extended now to a neighborhood $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ of $\boldsymbol{\theta}^\star$. Fix $\epsilon > 0$, independent of N, p, and q, and define

$$\begin{aligned}\widetilde{\lambda}^\epsilon_{\max,W} &:= \sup_{\boldsymbol{\theta}\in\mathcal{B}_2(\boldsymbol{\theta}^\star,\epsilon)} \frac{\lambda_{\max}\left(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell(\boldsymbol{\theta}, \boldsymbol{X})\right)}{K} \\ \widetilde{\lambda}^\epsilon_{\max,B} &:= \sup_{\boldsymbol{\theta}\in\mathcal{B}_2(\boldsymbol{\theta}^\star,\epsilon)} \frac{\lambda_{\max}\left(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_B}\,\ell(\boldsymbol{\theta}, \boldsymbol{X})\right)}{\binom{K}{2}}.\end{aligned} \tag{7}$$

We first establish lower bounds to the minimax risks $\mathscr{R}_{W,N}$ and $\mathscr{R}_{B,N}$ in Theorem 2.2, which enable us to outline sufficient conditions for the upper bounds on the ℓ_2-error presented in Theorem 2.1 to achieve (up to an unknown constant) the minimax rates of convergence; see Corollary 2.4. In the following results, it is helpful to recall that $\widetilde{\lambda}^\epsilon_{\min,W}$ and $\widetilde{\lambda}^\epsilon_{\min,B}$ are defined in Assumption 2, $\widetilde{\lambda}^\star_{\max,W}$ and $\widetilde{\lambda}^\star_{\max,B}$ are defined in Assumption 3, and $\widetilde{\lambda}^\epsilon_{\max,W}$ and $\widetilde{\lambda}^\epsilon_{\max,B}$ are defined in (7).

Theorem 2.2. *(Lower bound to the minimax risk) Consider an exponential-family local dependence random graph model satisfying Assumption 2. Then there exist constants $C_1 > 0$ and $C_2 > 0$, independent of N, p, and q, such that the minimax risks $\mathscr{R}_{W,N}$ and $\mathscr{R}_{B,N}$ defined in (6) satisfy*

$$\mathscr{R}_{W,N} \;\geq\; C_1 \sqrt{\frac{A_{\text{avg}}}{\widetilde{\lambda}^{\epsilon}_{\max,W}}} \sqrt{\frac{p}{N}} \;\geq\; C_1 \left(\frac{\widetilde{\lambda}^{\epsilon}_{\min,W}}{\widetilde{\lambda}^{\epsilon}_{\max,W}}\right) \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{A_{\text{avg}}} \sqrt{\frac{p}{N}}$$

$$\mathscr{R}_{B,N} \;\geq\; C_2 \frac{A_{\text{avg}}}{\sqrt{\widetilde{\lambda}^{\epsilon}_{\max,B}}} \sqrt{\frac{q}{N^2}} \;\geq\; C_2 \left(\frac{\widetilde{\lambda}^{\epsilon}_{\min,B}}{\widetilde{\lambda}^{\epsilon}_{\max,B}}\right) \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}} A_{\text{avg}} \sqrt{\frac{q}{N^2}},$$

provided $p = \dim(\boldsymbol{\theta}^{\star}_W) = O(N\, \widetilde{\lambda}^{\epsilon}_{\max,W})$ and $q = \dim(\boldsymbol{\theta}^{\star}_B) = O(N^2\, \widetilde{\lambda}^{\epsilon}_{\max,B})$.

The role of Assumption 2 in Theorem 2.2 is to ensure that both $\widetilde{\lambda}^{\epsilon}_{\min,W}$ and $\widetilde{\lambda}^{\epsilon}_{\min,B}$ are bounded away from 0, ensuring all of the lower bounds are well defined, whereas we assume

$$p \;=\; \dim(\boldsymbol{\theta}^{\star}_W) \;=\; O(N\, \widetilde{\lambda}^{\epsilon}_{\max,W}) \qquad \text{and} \qquad q \;=\; \dim(\boldsymbol{\theta}^{\star}_B) \;=\; O(N^2\, \widetilde{\lambda}^{\epsilon}_{\max,B}) \tag{8}$$

in order to satisfy a technical condition in the proof of Theorem 2.2. Under the assumption that the maximum eigenvalues of Fisher information matrices are bounded away from 0, the condition in (8) requires that $p = O(N)$ and $q = O(N^2)$, which places a much less stringent restriction on the dimensions of parameters vectors when compared with Assumption 3. Two sets of lower bounds are presented in Theorem 2.2, with the first being the most sharp, but unhelpful in our pursuit of studying whether the rates of convergence implied in Theorem 2.1 are minimax optimal. The second, though looser, set of bounds approximately match the upper bounds on the ℓ_2-error established in Theorem 2.1. Indeed, this second set of bounds allows us to establish conditions for such minimax optimality in Corollary 2.4 which is presented below.

Note that the lower bound to the minimax risk presented in Theorem 2.2 considers $\boldsymbol{\theta}^{\star} \in \mathbb{R}^{p+q}$. The fact that the parameter space is unbounded introduces no complications when deriving lower bounds; however, when turning to the problem of deriving an upper bound to the minimax risk, an unbounded parameter space presents new challenges. The following theorem obtains upper bounds on the minimax risk with respect to the ℓ_2-norm in a neighborhood of the data-generating parameter vector.

Theorem 2.3. *(Upper bound to the minimax risk) Under the assumptions of Theorem 2.1, there exist constants $C_1 > 0$, $C_2 > 0$, and $N_0 \geq 3$, independent of N, p, and q, such that the minimax risks restricted to a local neighborhood $\mathcal{B}_2(\boldsymbol{\theta}^{\star}, \epsilon)$ of a point $\boldsymbol{\theta}^{\star} \in \mathbb{R}^{p+q}$ satisfy, for all integers $N \geq N_0$,*

$$\inf_{\widehat{\boldsymbol{\theta}}_W} \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^{\star}, \epsilon)} \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \;\leq\; C_1 \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{A_{\text{avg}}} \sqrt{\frac{p}{N}}$$

$$\inf_{\widehat{\boldsymbol{\theta}}_B} \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^{\star}, \epsilon)} \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B\|_2 \;\leq\; C_2 \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}} A_{\text{avg}} \sqrt{\frac{q}{N^2}},$$

where $\epsilon > 0$ is the same as in Assumptions 2 and 3 and in (7).

The final result of this section is concerned with outlining a set of sufficient conditions which allow us to establish the minimax optimality of Theorem 2.1.

Corollary 2.4. *Under the assumptions of Theorems 2.1 and 2.2, and the assumption that*

$$\widetilde{\lambda}_{\max,W}^{\epsilon} = O\left(\widetilde{\lambda}_{\min,W}^{\epsilon}\right) \quad \text{and} \quad \widetilde{\lambda}_{\max,B}^{\epsilon} = O\left(\widetilde{\lambda}_{\min,B}^{\epsilon}\right), \tag{9}$$

the maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_W$ and $\widehat{\boldsymbol{\theta}}_B$ achieve the minimax rate of convergence, in the sense that the upper bounds on the ℓ_2-error of $\widehat{\boldsymbol{\theta}}_W$ and $\widehat{\boldsymbol{\theta}}_B$ presented in Theorem 2.1 match (up to an unknown constant which is independent of N, p, and q) the lower bounds to the minimax risks in Theorem 2.2.

If the exponential-family local dependence random graph model satisfies (9), then Corollary 2.4 establishes the minimax optimality of the rates of convergence for maximum likelihood estimators implied via Theorem 2.1. Such an assumption is common in the high-dimensional statistics literature (e.g., Ravikumar, Wainwright and Lafferty, 2010, Janková and van de Geer, 2018), where it is common to assume that minimum and maximum eigenvalues of Fisher information matrices corresponding to the sampling distribution are bounded away from 0 and from above, respectively. We can interpret condition (9) similarly, however applied to the joint Fisher information for the entire collection of random variables in the random graph (in contrast to the sampling distribution from which a random sample is generated) and in a neighborhood $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ of the data-generating parameter vector $\boldsymbol{\theta}^\star$.

2.3. Convergence rates of the multivariate normal approximation

A key challenge to any statistical analysis of network data is finding rigorous justification for statistical inference methodology. The main contributing factor to this challenge lies in the fact that statistical analyses of network data are typically in the setting of a single collection of dependent random variables without the benefit of replication. In other words, any statistical inference will be based on a single observation of a collection of dependent binary random variables. It is common for inference of model parameters in exponential-family random graph models to utilize the normal approximation for carrying out inference about estimated coefficients (e.g., Krivitsky et al., 2023, Lusher, Koskinen and Robins, 2012, Stewart et al., 2019). Except in select cases, these inferences are performed without rigorous theoretical justification, owing to the difficulty of obtaining theoretical results establishing the validity of the normal approximation in scenarios with a single observations of a collection of dependent binary random variables.

The dependence structure of local dependence random graph models facilitates proof of rigorous theoretical results justifying the normal approximation for estimators, and in this section, we obtain rates of convergence of the multivariate normal approximation in scenarios of increasing model dimension. It is worth noting that our results imply the univariate normal approximation, as multiple univariate tests are frequently utilized in applications (e.g., Stewart et al., 2019). Similarly to our consistency results presented in Theorem 2.1, the quality of the multivariate normal approximation will depend on key quantities related to the block structure, graph, and model specification.

Throughout, $\boldsymbol{Z}_d$ will denote a d-dimensional multivariate normal random vector with mean vector $\mathbf{0}_d$ (the d-dimensional vector of all zeros) and covariance matrix $\boldsymbol{I}_d$ (the d-dimensional identity matrix). The probability distribution of $\boldsymbol{Z}_d$ is denoted by Φ_d.

In order to establish our multivariate normal approximation theory, we leverage a multivariate Berry-Esseen theorem provided in Raič (2019), together with a Taylor expansion of the log-likelihood equation. Utilizing properties of exponential families, we are able to derive non-asymptotic bounds on the error of the multivariate normal approximation for a standardization of the maximum likelihood estimator, providing the first results which elaborate conditions under which the normal approximation is expected to produce valid inferences in local dependence random graph models.

Theorem 2.5. *Consider a minimal exponential-family local dependence random graph model satisfying Assumptions 1, 2, 3, and 4 and assume that* $p = \dim(\boldsymbol{\theta}_W^\star) \geq \log N$ *and* $q = \dim(\boldsymbol{\theta}_B^\star) \geq \log N$. *Then there exist constants* $C_1 > 0$, $C_2 > 0$, *and* $N_0 \geq 3$, *independent of* N, p, *and* q, *and a random vector* $\boldsymbol{\Delta} \in \mathbb{R}^{p+q}$ *such that, for all integers* $N \geq N_0$ *and measurable convex sets* $\mathscr{C} \subset \mathbb{R}^{p+q}$,

$$\left|\mathbb{P}(I(\boldsymbol{\theta}^\star)^{1/2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) + \boldsymbol{\Delta} \in \mathscr{C}) - \Phi_d(\mathbf{Z}_d \in \mathscr{C})\right|$$

$$\leq \quad C_1\,(p+q)^{1/4}\,A_{\max}^7 \left[\sqrt{\frac{p^3}{(\widetilde{\lambda}_{\min,W}^\star)^3\,N}} + \sqrt{\frac{q^3}{(\widetilde{\lambda}_{\min,B}^\star)^3\,N^2}}\,\right],$$

where the random vector $\boldsymbol{\Delta}$ *satisfies*

$$\mathbb{P}\left(\|\boldsymbol{\Delta}\|_2 \leq C_2\,A_{\max}^6 \sqrt{A_{\mathrm{avg}}\,\frac{(\widetilde{\lambda}_{\max,W}^\star)^2}{(\widetilde{\lambda}_{\min,W}^\epsilon)^5}\,\frac{p^5}{N} + A_{\mathrm{avg}}^2\,\frac{(\widetilde{\lambda}_{\max,B}^\star)^2}{(\widetilde{\lambda}_{\min,B}^\epsilon)^5}\,\frac{q^5}{N^2}}\;\right) \;\geq\; 1 - \frac{1}{N^2}.$$

The standardization $I(\boldsymbol{\theta}^\star)^{1/2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)$ is of a familiar form in multivariate normal approximation settings. The quantity $\boldsymbol{\Delta}$ can be interpreted as an error term or a random perturbation, arising due to a Taylor approximation. While our result is stated for $I(\boldsymbol{\theta}^\star)^{1/2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) + \boldsymbol{\Delta}$, an important aspect of Theorem 2.5 lies in establishing that the random perturbation $\boldsymbol{\Delta}$ to $I(\boldsymbol{\theta}^\star)^{1/2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)$ is small (in the ℓ_2-norm) with high probability, justifying basing inferences and derivations of confidence regions on $I(\boldsymbol{\theta}^\star)^{1/2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)$ in applications. Indeed, under mild assumptions (which we state below), it is straightforward to establish that $\|\boldsymbol{\Delta}\|_2$ converges almost surely to 0 as $N \to \infty$.

A remark is in order regarding the term $(p+q)^{1/4}$ in the upper bound on the error of the multivariate normal approximation in Theorem 2.5. Current results on multivariate Berry-Esseen bounds involve terms which are functions of the dimension of the random vector (Raič, 2019). Here, the total dimension of the random vector is $p+q$, as we are proving the joint multivariate normality of a standardization of the entire vector of maximum likelihood estimators $(\widehat{\boldsymbol{\theta}}_W, \widehat{\boldsymbol{\theta}}_B)$ which has dimension $p+q$. In other words, we are unable to separate the error into two terms which are functions of only quantities based on within-block and between-block quantities, as was done in our consistency theory in Section 2.2.

Typically, both $I_W(\boldsymbol{\theta}_W^\star)^{1/2}$ and $I_B(\boldsymbol{\theta}_B^\star)^{1/2}$ will be unknown, but can be approximated in practice. We can approximate both $I_W(\boldsymbol{\theta}_W^\star)$ and $I_B(\boldsymbol{\theta}_B^\star)$ through Monte-Carlo methods, as Fisher information matrices of canonical exponential families are the covariance matrices of the sufficient statistics. This is a common approach to estimating the Fisher information matrix in the exponential-family random graph model literature, owing to the fact that models are frequently estimated via Monte-Carlo maximum likelihood estimation, which already requires simulating sufficient statistic vectors (e.g., Hunter and Handcock, 2006, Krivitsky et al., 2023), and discussed in Section 2.1.

Under an additional regularity assumption, we can simplify the bounds presented in Theorem 2.5.

Assumption 5. Assume that there exist constants $L > 0$ and $U > 0$ such that

$$0 \;<\; L \;\leq\; \min\left\{\widetilde{\lambda}_{\min,W}^\epsilon,\, \widetilde{\lambda}_{\min,B}^\epsilon\right\} \;\leq\; \max\left\{\widetilde{\lambda}_{\max,W}^\star,\, \widetilde{\lambda}_{\max,B}^\star\right\} \;\leq\; U, \tag{10}$$

for all values of N, p, and q.

Assumption 5 is reminiscent of minimum and maximum eigenvalue restrictions in the high-dimensional statistics literature, where it is common to assume the minimum and maximum eigenvalues of Fisher information matrices are bounded away from 0 and from above, respectively (e.g., Ravikumar, Wainwright and Lafferty, 2010, Janková and van de Geer, 2018). Assumption 5 can be interpreted similarly, though applied to the averaged minimum and maximum eigenvalues of the joint Fisher information matrices; see also the discussions following Corollary 2.4.

Under Assumptions 1, 2, 3, 4, and 5, we may leverage Theorem 2.5 to establish, for all measurable convex sets $\mathscr{C} \subset \mathbb{R}^{p+q}$, the new bound of

$$|\mathbb{P}(I(\boldsymbol{\theta}^\star)^{1/2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) + \boldsymbol{\Delta} \in \mathscr{C}) - \Phi_d(\mathbf{Z}_d \in \mathscr{C})| \;\; \leq \;\; C_1\,(p+q)^{1/4}\, A_{\max}^7 \left[\sqrt{\frac{p^3}{N}} + \sqrt{\frac{q^3}{N^2}}\right],$$

where $\boldsymbol{\Delta}$ now satisfies

$$\mathbb{P}\left(\|\boldsymbol{\Delta}\|_2 \;\leq\; C_2\, A_{\max}^6 \sqrt{A_{\text{avg}}\,\frac{p^5}{N} + A_{\text{avg}}^2\,\frac{q^5}{N^2}}\right) \;\; \geq \;\; 1 - \frac{1}{N^2}.$$

In certain settings, it may be the case that properties of the network limit the sizes of the blocks, in which the size of the largest block $A_{\max}$ may be bounded for all network sizes. Under the additional assumption that the sizes of the blocks are bounded above, we can absorb the quantities involving $A_{\max}$ and A_{avg} into the constants $C_1 > 0$ and $C_2 > 0$ in the above bounds. This results in the following simple bounds on the error of the multivariate normal approximation:

$$\left|\mathbb{P}(I(\boldsymbol{\theta}^\star)^{1/2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) + \boldsymbol{\Delta} \in \mathscr{C}) - \Phi_d(\mathbf{Z}_d \in \mathscr{C})\right| \;\; \leq \;\; C_1\,(p+q)^{1/4} \left[\sqrt{\frac{p^3}{N}} + \sqrt{\frac{q^3}{N^2}}\right],$$

where $\boldsymbol{\Delta}$ will then satisfy

$$\mathbb{P}\left(\|\boldsymbol{\Delta}\|_2 \;\leq\; C_2 \sqrt{\frac{p^5}{N} + \frac{q^5}{N^2}}\right) \;\; \geq \;\; 1 - \frac{1}{N^2},$$

for all measurable convex sets $\mathscr{C} \subset \mathbb{R}^{p+q}$. Note, in the above results, that the probability bounds approach 1 sufficiently fast, allowing us to establish, through the Borel–Cantelli lemma, that $\|\boldsymbol{\Delta}\|_2$ converges $\mathbb{P}$-almost surely to 0 as $N \to \infty$, provided the upper bounds on $\|\boldsymbol{\Delta}\|_2$ tend to 0 as $N \to \infty$.

Finally, to deliver a simple and easily interpretable result for statistical inference, we prove a corollary to Theorem 2.5 establishing the asymptotic multivariate normality of maximum likelihood estimators.

Corollary 2.6. *Under the assumptions of Theorem 2.5, Assumption 5, and assuming*

$$\lim_{N\to\infty}\; \max\left\{A_{\max}^6 \sqrt{A_{\text{avg}}\,\frac{p^5}{N} + A_{\text{avg}}^2\,\frac{q^5}{N^2}},\;\; (p+q)^{1/4}\, A_{\max}^7 \left[\sqrt{\frac{p^3}{N}} + \sqrt{\frac{q^3}{N^2}}\right]\right\} \;\; = \;\; 0,$$

we have the distributional limit $I(\boldsymbol{\theta}^\star)^{1/2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) \xrightarrow{D} \mathbf{Z}_{p+q}$ *as* $N \to \infty$.

Corollary 2.6 can be proved directly by observing that the assumptions of the corollary ensure the error bounds in Theorem 2.5 converge to 0 in the limit as $N \to \infty$. As a result of Corollary 2.6, standard procedures for constructing confidence regions, univariate confidence intervals, and performing statistical hypothesis tests for significance of parameters are justified using the asymptotic approximation of the variance-covariance matrix $I(\boldsymbol{\theta}^\star) = \mathbb{V}\, s(\boldsymbol{X})$, which we discuss above. When the sizes of the blocks are bounded as above, the essential condition for asymptotic multivariate normality because

$$\lim_{N \to \infty} \sqrt{\frac{p^5}{N} + \frac{q^5}{N^2}} \;\; = \;\; 0,$$

restricting the maximum growth with N of the dimensions of the parameters vectors $p = \dim(\boldsymbol{\theta}_W)$ and $q = \dim(\boldsymbol{\theta}_B)$, suggesting that both $p = \dim(\boldsymbol{\theta}_W) = o(N^{1/5})$ and $q = \dim(\boldsymbol{\theta}_B) = o(N^{2/5})$ must hold in our theory for the error of the multivariate normal approximation to vanish in the limit as $N \to \infty$.

3. Simulation results

3.1. Simulation study 1: Convergence rates of maximum likelihood estimators

Simulation study 1 demonstrates that the rate of growth of the dimension of parameter vectors plays a key role in the finite sample performance. We consider three cases in a setting which controls certain aspects of the graph. Throughout this study, we assume that the sizes of the blocks are all fixed at 50, i.e., $|\mathcal{A}_k| = 50$ for all $k \in \{1, \ldots, K\}$. In order to vary the size of the network N, we vary the number of blocks $K \in \{1, 5, 10, 15, 20\}$, which results in networks of size $N \in \{50, 250, 500, 750, 1000\}$. We focus on a special case of Example 3 from Section 1.2, by assuming that each node $i \in \mathcal{N}$ is assigned to a group $\mathcal{G}_1, \ldots, \mathcal{G}_M$ ($M \geq 2$). The specific form of this model is then given by

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) \;\; \propto \;\; \exp\left(\sum_{m=1}^{M} \theta_m\, s_m(\boldsymbol{x}) + \theta_{m+1}\, s_{m+1}(\boldsymbol{x})\right),$$

where

$$s_m(\boldsymbol{x}) \;\; = \;\; \sum_{k=1}^{K} \; \sum_{i \in \mathcal{A}_k \cap \mathcal{G}_m} \; \sum_{j \in \mathcal{A}_k \setminus \{i\}} x_{i,j}, \qquad m \in \{1, \ldots, M\},$$

and

$$s_{m+1}(\boldsymbol{x}) \;\; = \;\; \sum_{k=1}^{K} \; \sum_{i<j \,:\, i \in \mathcal{A}_k,\, j \in \mathcal{A}_k} x_{i,j}\, \mathbb{1}\left(\sum_{h \in \mathcal{A}_k \setminus \{i,j\}} x_{i,h}\, x_{j,h} \;\geq\; 1\right).$$

For this simulation study we will focus on the within-block parameter vector in order to easily compare the trade-off between the dimension of the parameter vector p and the size of the network N. We can then assume that $X_{i,j} = 0$ with probability one for all $\{i, j\} \subset \mathcal{N}$ belonging to distinct blocks, i.e., the between-block subgraphs $\boldsymbol{X}_{k,l}$ ($1 \leq k < l \leq K$) are empty subgraphs with probability one.

We consider three cases:

- Case 1: $M = 3$, in which case $p = 4$ for all $N \in \{50, 250, 500, 750, 1000\}$.
- Case 2: $M = \lceil N^{2/5} \rceil$, in which case $p \in \{6, 11, 14, 16, 17\}$ depending on the size of the network.
- Case 3: $M = \lceil \sqrt{N} \rceil$, in which case $p \in \{9, 17, 24, 29, 33\}$ depending on the size of the network.

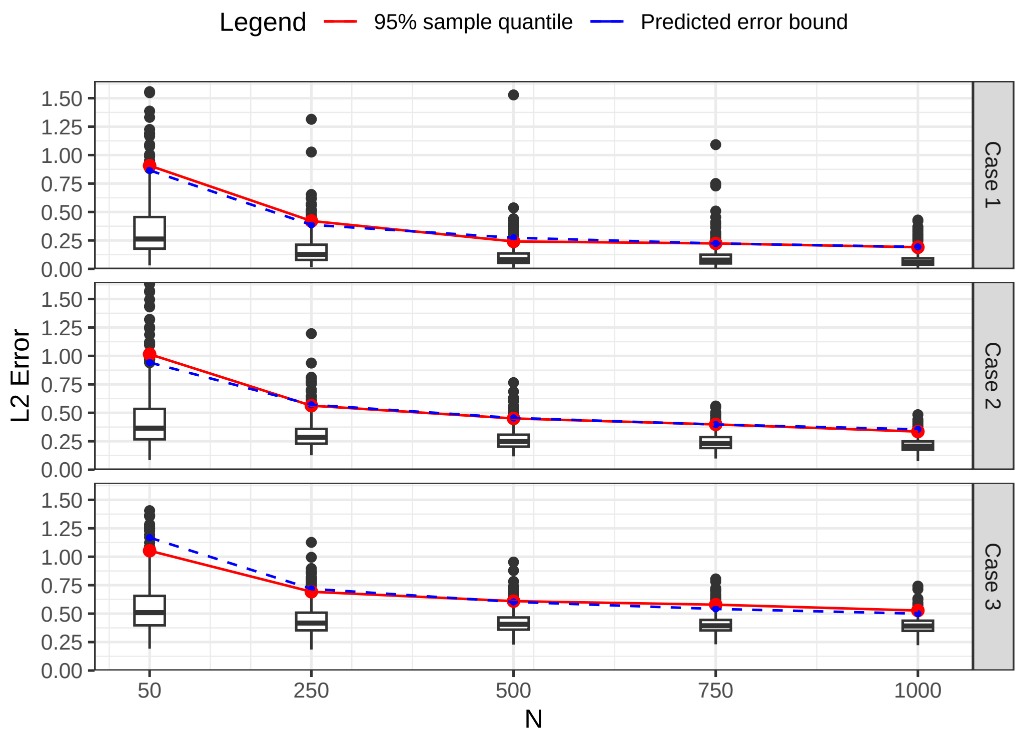

Figure 2. The results of Simulation study 1, which demonstrates the trade-off in finite sample performance of maximum likelihood estimators based on the number of model parameters and size of the network. Each boxplot for each combination of case and network size is based on 500 replications. Boxplots display the empirical distribution of the ℓ_2-error, whereas the red lines track the 95% sample quantiles and the blue dashed lines track the error bounds predicted by Theorem 2.1.

For each case and network size $N \in \{50, 250, 500, 750, 1000\}$, we simulate 500 networks from $\mathbb{P}_{\boldsymbol{\theta}}$ where $\theta_{M+1} = .5$ and $(\theta_1, \ldots, \theta_M) \overset{iid}{\sim} \text{Unif}(-1.5, -.5)$. The value of θ_{M+1} ensures there is a reasonably strong tendency towards transitivity in the network, and the values of $(\theta_1, \ldots, \theta_M)$ result in networks with plausible densities. The results of the Simulation study 1 are summarized in Figure 2.

The finite sample performance of this study suggests, as would be expected based on the results of Theorem 2.1, that the rate at which the ℓ_2-error converges to 0 is fastest in Case 1 for which the model dimension is fixed, and slowest in Case 3 for which the model dimension is on the order of $\sqrt{N}$. We compute a predicted error bound based on Theorem 2.1 by estimating the constant terms, which in this simulation study include the average block sizes A_{avg} and the largest block size $A_{\max}$, as well as the terms quantifying averaged eigenvalues of the Fisher information matrices. This can be accomplished by estimating constants for each network size by

$$\widehat{C}_N \quad := \quad Q_{N,.95} \Bigg/ \sqrt{\frac{p}{N}}, \qquad\qquad N \in \{50, 250, 500, 750, 1000\},$$

where $Q_{N,.95}$ is the 95% sample quantile of the ℓ_2-errors of the maximum likelihood estimators based on the 500 replications, and then using the estimate

$$\widehat{C} \;\coloneqq\; \frac{1}{5} \sum_{N \in \{50,250,500,750,1000\}} \widehat{C}_N$$

to obtain an overall estimate of the constant term. The predicted error bounds are then defined as

$$\widetilde{E}_N \;\coloneqq\; \widehat{C} \sqrt{\frac{p}{N}}, \qquad N \in \{50, 250, 500, 750, 1000\}.$$

The dashed blue lines track the values of $\widetilde{E}_N$ in Figure 2, whereas the red lines track $Q_{N,.95}$.

Notably, the predicted error bound closely matches the 95% sample quantile of the simulated ℓ_2-errors. Theorem 2.1 establishes a bound which should hold with high probability, provided N is sufficiently large. Figure 2 demonstrates that the predicted error bounds most closely match the realized 95% sample quantile of the simulated ℓ_2-errors for larger network sizes. It is also worth noting that an additional source of variation here may be due to the fact that the constant term is not actually constant in the network size, as the quantities $\widetilde{\lambda}^{\star}_{\max,W}$ and $\widetilde{\lambda}^{\epsilon}_{\min,W}$ may depend on N. With that said, though, the simulation reveals close agreement with the predicted error bounds.

3.2. Simulation study 2: Error of the normal approximation

The second simulation study we conduct explores the error of the normal approximation, leveraging results in Theorem 2.5. We consider the same probability distribution as in Simulation study 1, in the following two cases:

- Case 1: Fixed parameter dimension $p = 5$ with $M = 4$ categories of each node group and networks of size $N \in \{250, 500, 750, 1000\}$.
- Case 2: Growing parameter dimension $p = 2K$ with $M = 2K - 1$ categories of each node group, where there are 50 nodes per block and the number of blocks vary over $K \in \{5, 10, 15, 20\}$, resulting in networks of size $N \in \{250, 500, 750, 1000\}$.

We generate 500 replications in each case, simulating networks from the same probability distributions as in Simulation study 1 and in the same manner.

We study the quality of the normal approximation by constructing confidence intervals for the transitive edge parameter and Quantile-Quantile plots for the standardized maximum likelihood estimator of the transitive edge parameter. Our results demonstrate the empirical Type I error in the former matches the theoretical Type I error, with the Quantile-Quantile plots not revealing significant departure from normality. For each case, we constructed 95% confidence intervals and computed the empirical Type I error control. Letting θ_{m+1} and $\widehat{\theta}_{m+1}$ denote the transitive edge parameter and the maximum likelihood estimator of the transitive edge parameter, we leverage Theorem 2.5 to construct confidence intervals:

$$\mathbb{P}\left(\theta^{\star}_{m+1} \in \left[\widehat{\theta}_{m+1} - q_{1-\alpha/2}\sqrt{[\boldsymbol{S}^{-1}]_{m+1,m+1}},\; \widehat{\theta}_{m+1} + q_{1-\alpha/2}\sqrt{[\boldsymbol{S}^{-1}]_{m+1,m+1}}\right]\right) \approx 1 - \alpha, \quad \alpha \in (0,1),$$

where $q_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$%-quantile of the univariate standard normal distribution and $\boldsymbol{S}$ denotes the sample variance-covariance matrix obtained by sampling sufficient statistics through MCMC methods; see the discussions in Section 2.1. For Case 1, the empirical coverage was $(.96, .95, .95, .96)$ corresponding to network sizes of $(250, 500, 750, 1000)$, and for Case 2, the same was $(.96, .95, .95, .96)$ corresponding to network sizes of $(250, 500, 750, 1000)$. The Quantile-Quantile plots for each case across the different network sizes are presented in Figure 3.

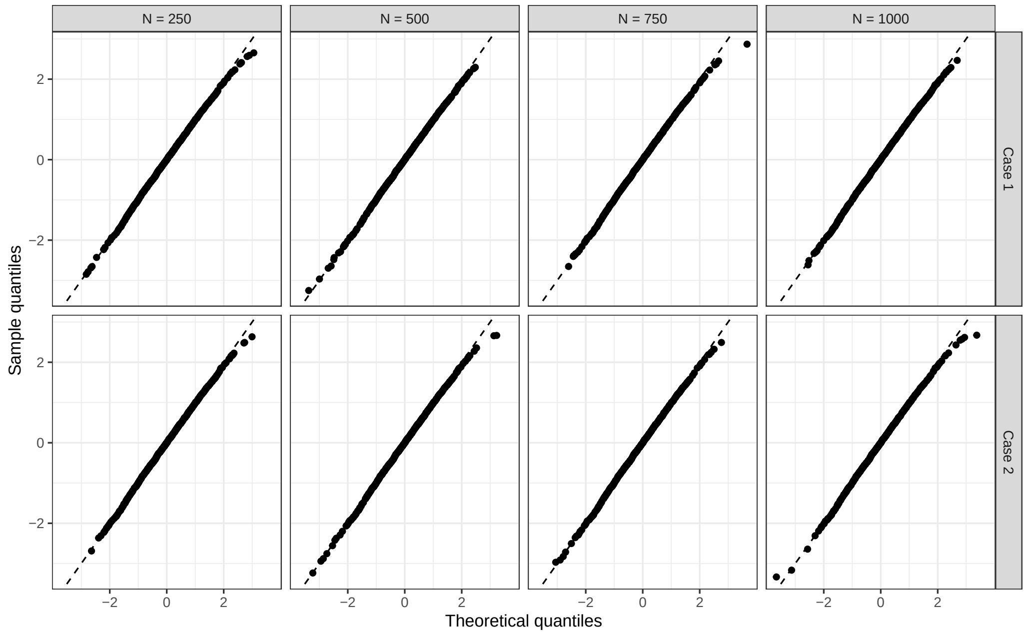

Figure 3. Quantile-Quantile plots showing the results of Simulation Study 2. The sample quantiles of the standardized maximum likelihood estimates of the transitive edge parameter are plotted against the theoretical quantiles based on the standard normal approximation in each of the two cases studied in Simulation study 2 across networks of size $N \in \{250, 500, 750, 1000\}$.

4. Conclusions

In this work, we have proved the first rigorous theory for both estimation and statistical inference of local dependence random graph models. We have established minimax optimal rates of convergence in the ℓ_2-norm of maximum likelihood estimators of exponential-family local dependence random graph models, accompanying these results with finite-sample error bounds on the multivariate normal approximation of a standardization of maximum likelihood estimators. Notably, our results allow for both the number of parameters and the sizes of blocks to grow unbounded with the size of the network.

Our consistency and normal approximation theory are non-asymptotic, although we have stated helpful asymptotic results along the way, which enable us to understand how key aspects of the model (through the spectrum of Fisher information matrices and the dimension of parameter vectors) and properties of the network (through the number and sizes of blocks and nodes) impact rates of convergence for both the statistical error (in the ℓ_2-norm) and the multivariate normal approximation. Our results cover general settings and heterogeneous parameterizations, as exemplified in the examples in Section 1.2 and our simulation studies in Section 3, which allow our results to cover a broad scope.

Results were derived under the assumption that we have perfect knowledge of the block memberships of nodes in the network. This may be reasonable in certain settings where we can observe the block memberships of nodes, but might be violated in other settings where we obtain imperfect observations of the block memberships of nodes, whether through a noisy observation process or error in the estimates of the block memberships. The effect of imperfect knowledge of the block memberships of nodes on the aforementioned errors and convergence rates is an open question.

Acknowledgments

The author is grateful to two anonymous referees, an associate editor, the editor, and to Dr. Michael Schweinberger, for constructive comments and suggestions that greatly improved the paper.

Funding

The author was supported by NSF award MMS-2345043 and the Department of Defense Test Resource Management Center under contracts FA8075-18-D-0002 and FA8075-21-F-0074.

Supplementary Material

Supplement to "Rates of convergence and normal approximations for estimators of local dependence random graph models". The supplementary material contains the proofs for all results in Section 2, as well as additional technical results and the proofs of these technical results used to prove the main results in Section 2.

References

AGNEESSENS, F., TRINCADO-MUNOZ, F. J. and KOSKINEN, J. (2024). Network formation in organizational settings: Exploring the importance of local social processes and team-level contextual variables in small groups using bayesian hierarchical ERGMs. *Soc. Netw.* **77** 104–117.

BABKIN, S., STEWART, J. R., LONG, X. and SCHWEINBERGER, M. (2020). Large-scale estimation of random graph models with local dependence. *Comput. Statist. Data Anal.* **152** 107029.

BROWN, L. D. (1986). *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics.

CHATTERJEE, S., DIACONIS, P. and SLY, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21** 1400–1435.

CHATTERJEE, S. and DIACONIS, P. (2013). Estimating and understanding exponential random graph models. *Ann. Statist.* **41** 2428–2461.

CHEN, P., GAO, C. and ZHANG, A. Y. (2022). Partial recovery for top-k ranking: optimality of MLE and suboptimality of spectral method. *Ann. Statist.* **50** 1618–1652.

DAHBURA, J. N. M., KOMATSU, S., NISHIDA, T. and MELE, A. (2021). A structural model of business card exchange networks. *arXiv preprint arXiv:2105.12704*.

DAHBURA, J. N. M., KOMATSU, S., NISHIDA, T. and MELE, A. (2023). Homophily and community structure at scale: an application to a large professional network. *AEA Papers and Proc.* **113** 156–160.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. R. Stat. Soc. Ser. B Stat. Method* **54** 657–683.

GHOSAL, S. (2000). Asymptotic Normality of Posterior Distributions for Exponential Families when the Number of Parameters Tends to Infinity. *J. Multivariate Anal.* **74** 49–68.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137.

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–50.

HUNTER, D. R. and HANDCOCK, M. S. (2006). Inference in curved exponential family models for networks. *J. Comput. Graph. Statist.* **15** 565–583.

HÄGGSTRÖM, O. and JONASSON, J. (1999). Phase transition in the random triangle model. *J. Appl. Probab.* **36** 1101–1115.

JANKOVÁ, J. and VAN DE GEER, S. (2018). Inference in high-dimensional graphical models. In *Handbook of Graphical Models* (M. Maathuis, M. Drton, S. Lauritzen and M. Wainwright, eds.) 325–350. CRC Press.

JONASSON, J. (1999). The random triangle model. *J. Appl. Probab.* **36** 852–867.

KRIVITSKY, P. N., HUNTER, D. R., MORRIS, M. and KLUMB, C. (2023). ergm 4: new features for analyzing exponential-family random graph models. *J. Stat. Softw.* **105** 1–44.

LUSHER, D., KOSKINEN, J. and ROBINS, G. (2012). *Exponential Random Graph Models for Soc. Netw.* Cambridge University Press.

MELE, A. (2022). A structural model of homophily and clustering in social networks. *J. Bus. Econom. Statist.* **40** 1377–1389.

NEWMAN, M. E. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69** 026113.

ORTEGA, J. M. and RHEINBOLDT, W. C. (2000). *Iterative solution of nonlinear equations in several variables*. Society for Industrial and Applied Mathematics.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366.

RAIČ, M. (2019). A multivariate Berry-Esseen theorem with explicit constants. *Bernoulli* **25** 2824–2853.

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319.

SCHUBERT, D. and BRAND, A. (2022). "Whom should I talk to?": role prescription and hierarchy building in supervised living groups. *Soc. Incl.* **10** 295–306.

SCHWEINBERGER, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *J. Am. Stat. Assoc.* **106** 1361–1370.

SCHWEINBERGER, M. (2020). Consistent structure estimation of exponential-family random graph models with block structure. *Bernoulli* **26** 1205–1233.

SCHWEINBERGER, M. and HANDCOCK, M. S. (2015). Local dependence in random graph models: characterization, properties and statistical inference. *J. R. Stat. Soc. Ser. B Stat. Method.* **77** 647–676.

SCHWEINBERGER, M. and STEWART, J. (2020). Concentration and consistency results for canonical and curved exponential-family models of random graphs. *Ann. Stat.* **48** 374–396.

SCHWEINBERGER, M., KRIVITSKY, P. N., BUTTS, C. T. and STEWART, J. R. (2020). Exponential-family models of random graphs: inference in finite, super and infinite population scenarios. *Stat. Sci.* **35** 627–662.

SNIJDERS, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *J. Soc. Struct.* **3** 1–40.

STEWART, J. R. (2024). Supplement to "Rates of convergence and normal approximations for estimators of local dependence random graph models".

STEWART, J. R. and SCHWEINBERGER, M. (2019). R package mlergm. Multilevel Exponential-Family Random Graph Models. *The Comprehensive R Archive Network*.

STEWART, J. R. and SCHWEINBERGER, M. (2020). Pseudo-likelihood-based M-estimation of random graphs with dependent edges and parameter vectors of increasing dimension. *arXiv preprint arXiv:2012.07167*.

STEWART, J., SCHWEINBERGER, M., BOJANOWSKI, M. and MORRIS, M. (2019). Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms. *Soc. Netw.* **59** 98–119.

SUNDBERG, R. (2019). *Statistical Modelling by Exponential Families*. Cambridge University Press.

TOLOCHKO, P. and BOOMGAARDEN, H. G. (2024). Same but different: a comparison of estimation approaches for exponential random graph models for multiple networks. *Soc. Netw.* **76** 1–11.

VERSHYNIN, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge university press.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge university press.

WHETSELL, T. A., KROLL, A. and DEHART-DAVIS, L. (2021). Formal hierarchies and informal networks: how organizational structure shapes information search in local government. *J. Public Adm. Res. Theory* **31** 653–669.

Supplement to “Rates of convergence and normal approximations for estimators of local dependence random graph models”

JONATHAN R. STEWART[1,a,a]

[1]*Department of Statistics, Florida State University, Tallahassee, FL, USA,* [a]*jrstewart@fsu.edu*

Section 1: Proof of Theorem 2.1 1

Section 1.1: Auxiliary results for Theorem 2.1 8

Section 2: Proof of Theorem 2.2 12

Section 3: Proof of Theorem 2.3 17

Section 4: Proof of Corollary 2.4 19

Section 5: Proof of Theorem 2.5 20

Section 5.1: Auxiliary results for Theorem 2.5 27

Section 6: Auxiliary results for exponential families 31

1. Proof of Theorem 2.1

Our method of proof utilizes a general M-estimation argument. For ease of presentation, we first introduce the general argument and then apply the general argument to maximum likelihood estimators of exponential-family local dependence random graph models to obtain the results of Theorem 2.1.

General M-estimation framework for rates of convergence. Consider a random estimating function $m : \mathbb{R}^d \times \mathbb{X} \mapsto \mathbb{R}$ and define $M(\boldsymbol{\theta}) := \mathbb{E}\, m(\boldsymbol{\theta}, \boldsymbol{X})$ for $\boldsymbol{\theta} \in \mathbb{R}^d$. We make the following assumptions concerning $m(\boldsymbol{\theta}, \boldsymbol{x})$ and $M(\boldsymbol{\theta})$:

1. Assume that $m(\boldsymbol{\theta}, \boldsymbol{x})$ is concave in $\boldsymbol{\theta} \in \mathbb{R}^d$ and continuously differentiable at all $\boldsymbol{\theta} \in \mathbb{R}^d$ and for all $\boldsymbol{x} \in \mathbb{X}$.
2. Assume that $M(\boldsymbol{\theta})$ is strictly concave in $\boldsymbol{\theta} \in \mathbb{R}^d$ and that $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is the unique global maximizer of $M(\boldsymbol{\theta})$.
3. Assume that $M(\boldsymbol{\theta})$ is twice continuously differentiable and that there exists an $\epsilon > 0$ (fixed) such that the negative Hessian $\boldsymbol{H}(\boldsymbol{\theta}) := -\nabla^2_{\boldsymbol{\theta}} M(\boldsymbol{\theta})$ of $M(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$.

When $m(\boldsymbol{\theta}, \boldsymbol{x})$ is the log-likelihood corresponding to a minimal exponential family, standard exponential family theory establishes that the above conditions (1) and (2) hold (e.g., Proposition 3.10 of

(Sundberg, 2019)). As a result, $\nabla_{\boldsymbol{\theta}}\, M(\boldsymbol{\theta}^\star) = \mathbf{0}_d$, where $\mathbf{0}_d$ is the d-dimensional zero vector. By Theorem 6.3.4 of Ortega and Rheinboldt (2000), if the event (for $\delta > 0$)

$$\inf_{\boldsymbol{\theta} \in \partial \mathcal{B}_2(\boldsymbol{\theta}^\star,\, \delta)} \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}, \boldsymbol{X}) \rangle \;\; \geq \;\; 0 \tag{11}$$

occurs, where $\partial \mathcal{B}_2(\boldsymbol{\theta}^\star,\, \delta)$ denotes the boundary of $\mathcal{B}_2(\boldsymbol{\theta}^\star,\, \delta) \coloneqq \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 < \delta\}$, then a root of $\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}, \boldsymbol{X})$ exists in $\overline{\mathcal{B}_2}(\boldsymbol{\theta}^\star, \delta) \coloneqq \mathcal{B}_2(\boldsymbol{\theta}^\star,\, \delta) \cup \partial \mathcal{B}_2(\boldsymbol{\theta}^\star,\, \delta)$, in which case a global maximizer $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^d}\, m(\boldsymbol{\theta}, \boldsymbol{X})$ exists and satisfies $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_2 \leq \delta$.

The key to our approach lies in demonstrating that condition (11) holds with high probability for a chosen $\delta \in (0, \epsilon)$ ($\epsilon > 0$ fixed) which helps to establish rates of convergence of estimators. In order to do so, we leverage the multivariate mean value theorem to establish that there exists, for each parameter vector $\boldsymbol{\theta} \in \partial \mathcal{B}_2(\boldsymbol{\theta}^\star,\, \delta)$, a parameter vector

$$\dot{\boldsymbol{\theta}} \;\; = \;\; t\, \boldsymbol{\theta} + (1-t)\, \boldsymbol{\theta}^\star \;\; \in \;\; \mathcal{B}_2(\boldsymbol{\theta}^\star,\, \delta) \;\; \subset \;\; \mathcal{B}_2(\boldsymbol{\theta}^\star,\, \epsilon), \qquad \text{for some } t \in (0, 1),$$

such that

$$\begin{aligned} \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \nabla_{\boldsymbol{\theta}}\, M(\boldsymbol{\theta}) \rangle \;\; &= \;\; \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \nabla_{\boldsymbol{\theta}}\, M(\boldsymbol{\theta}^\star) \rangle + \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \boldsymbol{H}(\dot{\boldsymbol{\theta}})\, (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \rangle \\ &= \;\; \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \boldsymbol{H}(\dot{\boldsymbol{\theta}})\, (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \rangle, \end{aligned}$$

recalling that $\nabla_{\boldsymbol{\theta}}\, M(\boldsymbol{\theta}^\star) = \mathbf{0}_d$. Observe that

$$\langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \boldsymbol{H}(\dot{\boldsymbol{\theta}})\, (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \rangle \;\; = \;\; \frac{\langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \boldsymbol{H}(\dot{\boldsymbol{\theta}})\, (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \rangle}{\langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle}\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \;\; \geq \;\; \lambda_{\min}(\boldsymbol{H}(\dot{\boldsymbol{\theta}}))\, \delta^2,$$

noting that the Rayleigh quotient of $\boldsymbol{H}(\dot{\boldsymbol{\theta}})$ is bounded below by the smallest eigenvalue $\lambda_{\min}(\boldsymbol{H}(\dot{\boldsymbol{\theta}}))$ of $\boldsymbol{H}(\dot{\boldsymbol{\theta}})$ and that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 = \delta$ for all $\boldsymbol{\theta} \in \partial \mathcal{B}_2(\boldsymbol{\theta}^\star, \delta)$. As a result,

$$\inf_{\boldsymbol{\theta} \in \partial \mathcal{B}_2(\boldsymbol{\theta}^\star, \delta)} \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \nabla_{\boldsymbol{\theta}}\, M(\boldsymbol{\theta}) \rangle \;\; \geq \;\; \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \delta)} \lambda_{\min}(\boldsymbol{H}(\boldsymbol{\theta}))\, \delta^2. \tag{12}$$

As $\partial \mathcal{B}_2(\boldsymbol{\theta}^\star, \delta) \subset \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$,

$$\inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \delta)} \lambda_{\min}(\boldsymbol{H}(\boldsymbol{\theta})) \;\; \geq \;\; \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \lambda_{\min}(\boldsymbol{H}(\boldsymbol{\theta})) \;\; > \;\; 0, \tag{13}$$

by the assumption that $\boldsymbol{H}(\boldsymbol{\theta})$ is positive definite, and thus non-singular, on $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$; Assumption 2 ensures this condition for maximum likelihood estimators of exponential-family local dependence random graph models. As a result of (12) and (13), the event

$$\sup_{\boldsymbol{\theta} \in \partial \mathcal{B}_2(\boldsymbol{\theta}^\star, \delta)} \left| \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star,\, \nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}, \boldsymbol{X}) - \nabla_{\boldsymbol{\theta}}\, M(\boldsymbol{\theta}) \rangle \right| \leq \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \lambda_{\min}(\boldsymbol{H}(\boldsymbol{\theta}))\, \delta^2 \tag{14}$$

implies the event (11). Thus, demonstrating that event (14) occurs with probability at least $1 - N^{-2}$ demonstrates that event (11) occurs with probability at least $1 - N^{-2}$.

Rates of convergence for maximum likelihood estimators. The log-likelihood equation of an exponential-family local dependence random graph model has the form

$$\ell(\boldsymbol{\theta}, \boldsymbol{x}) \;\; = \;\; \sum_{k=1}^{K} \ell_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) + \sum_{1 \leq k < l \leq K} \ell_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}),$$

which implies that the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_W, \widehat{\boldsymbol{\theta}}_B)$ is given by

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_W &= \operatorname*{arg\,max}_{\boldsymbol{\theta}_W \in \mathbb{R}^p} \sum_{k=1}^{K} \ell_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) \\ \widehat{\boldsymbol{\theta}}_B &= \operatorname*{arg\,max}_{\boldsymbol{\theta}_B \in \mathbb{R}^q} \sum_{1 \le k < l \le K} \ell_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}),\end{aligned} \tag{15}$$

owing to the fact that the subgraphs $\boldsymbol{X}_{k,l}$ $(1 \le k \le l \le K)$ are independent and that the parameter vectors $\boldsymbol{\theta}_W$ and $\boldsymbol{\theta}_B$ partition the parameters in $\boldsymbol{\theta}$. Hence, each optimizer in (15) can be found separately and independently. Define

$$\begin{aligned}m_W(\boldsymbol{\theta}_W, \boldsymbol{x}_W) &\coloneqq \sum_{k=1}^{K} \ell_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) \\ m_B(\boldsymbol{\theta}_B, \boldsymbol{x}_B) &\coloneqq \sum_{1 \le k < l \le K} \ell_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}),\end{aligned}$$

$M_W(\boldsymbol{\theta}_W) \coloneqq \mathbb{E}\, m_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W)$, and $M_B(\boldsymbol{\theta}_B) \coloneqq \mathbb{E}\, m_B(\boldsymbol{\theta}_B, \boldsymbol{X}_B)$, where

$$\boldsymbol{X}_W \coloneqq (\boldsymbol{X}_{1,1}, \ldots, \boldsymbol{X}_{K,K}) \quad \text{and} \quad \boldsymbol{X}_B \coloneqq (\boldsymbol{X}_{1,2}, \ldots, \boldsymbol{X}_{1,K}, \boldsymbol{X}_{2,3}, \boldsymbol{X}_{2,4}, \ldots, \boldsymbol{X}_{K-1,K}).$$

Due to the above considerations,

$$\boldsymbol{H}(\boldsymbol{\theta}) \; \coloneqq \; -\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{X}) \; = \; \begin{pmatrix} \boldsymbol{H}_W(\boldsymbol{\theta}_W) & \boldsymbol{0}_{p,q} \\ \boldsymbol{0}_{q,p} & \boldsymbol{H}_B(\boldsymbol{\theta}_B) \end{pmatrix},$$

where $\boldsymbol{0}_{d,r}$ is the $(d \times r)$-dimensional matrix of all zeros, and where

$$\boldsymbol{H}_W(\boldsymbol{\theta}_W) \coloneqq \sum_{k=1}^{K} \boldsymbol{H}_{k,k}(\boldsymbol{\theta}_W) \quad \text{and} \quad \boldsymbol{H}_B(\boldsymbol{\theta}_B) \coloneqq \sum_{1 \le k < l \le K} \boldsymbol{H}_{k,l}(\boldsymbol{\theta}_B),$$

with the definitions

$$\begin{aligned}\boldsymbol{H}_{k,k}(\boldsymbol{\theta}_W) &\coloneqq -\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_W}\, \ell_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{X}_{k,k}), && \text{for all } 1 \le k \le K \\ \boldsymbol{H}_{k,l}(\boldsymbol{\theta}_B) &\coloneqq -\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_B}\, \ell_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{X}_{k,l}), && \text{for all } 1 \le k < l \le K.\end{aligned}$$

Note that the interchange of differentiation and integration in this setting is trivial as the expectations are finite sums.

We demonstrate that event (14) occurs with probability at least $1 - N^{-2}$ for the within-block and between-block cases separately. Assumption 2 ensures that

$$\begin{aligned}\inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star_W, \epsilon)} \lambda_{\min}(\boldsymbol{H}_W(\boldsymbol{\theta}_W)) &= K\, \widetilde{\lambda}^{\epsilon}_{\min,W} > 0 \\ \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star_B, \epsilon)} \lambda_{\min}(\boldsymbol{H}_B(\boldsymbol{\theta}_B)) &= \binom{K}{2} \widetilde{\lambda}^{\epsilon}_{\min,B} > 0.\end{aligned} \tag{16}$$

Let $\delta_W \in (0, \epsilon\,/\,\sqrt{2})$ and $\delta_B \in (0, \epsilon\,/\,\sqrt{2})$, and assume $\boldsymbol{\theta}_W \in \partial\mathcal{B}_2(\boldsymbol{\theta}_W^\star, \delta_W)$ and $\boldsymbol{\theta}_B \in \partial\mathcal{B}_2(\boldsymbol{\theta}_B^\star, \delta_B)$. By assumption, $(\boldsymbol{\theta}_W, \boldsymbol{\theta}_B) \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$. Thus, using (16), we can rewrite the events in (14) as events

$$\begin{aligned}
|\langle \boldsymbol{\theta}_W - \boldsymbol{\theta}_W^\star, \nabla_{\boldsymbol{\theta}_W} m_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W) - \nabla_{\boldsymbol{\theta}_W} M_W(\boldsymbol{\theta}_W)\rangle| \;&\leq\; \delta_W^2\, K\, \widetilde{\lambda}_{\min,W}^{\epsilon} \\
|\langle \boldsymbol{\theta}_B - \boldsymbol{\theta}_B^\star, \nabla_{\boldsymbol{\theta}_B} m_B(\boldsymbol{\theta}_B, \boldsymbol{X}_B) - \nabla_{\boldsymbol{\theta}_B} M_B(\boldsymbol{\theta}_B)\rangle| \;&\leq\; \delta_B^2 \binom{K}{2} \widetilde{\lambda}_{\min,B}^{\epsilon}.
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&|\langle \boldsymbol{\theta}_W - \boldsymbol{\theta}_W^\star, \nabla_{\boldsymbol{\theta}_W} m_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W) - \nabla_{\boldsymbol{\theta}_W} M_W(\boldsymbol{\theta}_W)\rangle| \\
\leq\;& \|\boldsymbol{\theta}_W - \boldsymbol{\theta}_W^\star\|_2\, \|\nabla_{\boldsymbol{\theta}_W} m_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W) - \nabla_{\boldsymbol{\theta}_W} M_W(\boldsymbol{\theta}_W)\|_2 \\
=\;& \delta_W\, \|\nabla_{\boldsymbol{\theta}_W} m_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W) - \nabla_{\boldsymbol{\theta}_W} M_W(\boldsymbol{\theta}_W)\|_2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&|\langle \boldsymbol{\theta}_B - \boldsymbol{\theta}_B^\star, \nabla_{\boldsymbol{\theta}_B} m_B(\boldsymbol{\theta}_B, \boldsymbol{X}_B) - \nabla_{\boldsymbol{\theta}_B} M_B(\boldsymbol{\theta}_B)\rangle| \\
\leq\;& \delta_B\, \|\nabla_{\boldsymbol{\theta}_B} m_B(\boldsymbol{\theta}_B, \boldsymbol{X}_B) - \nabla_{\boldsymbol{\theta}_B} M_B(\boldsymbol{\theta}_B)\|_2.
\end{aligned}$$

It therefore suffices to demonstrate, for all $\boldsymbol{\theta}_W \in \partial\mathcal{B}_2(\boldsymbol{\theta}_W^\star, \delta_W)$ and $\boldsymbol{\theta}_B \in \partial\mathcal{B}_2(\boldsymbol{\theta}_B^\star, \delta_B)$, that events

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}_W} m_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W) - \nabla_{\boldsymbol{\theta}_W} M_W(\boldsymbol{\theta}_W)\|_2 \;&\leq\; \delta_W\, K\, \widetilde{\lambda}_{\min,W}^{\epsilon} \\
\|\nabla_{\boldsymbol{\theta}_B} m_B(\boldsymbol{\theta}_B, \boldsymbol{X}_B) - \nabla_{\boldsymbol{\theta}_B} M_B(\boldsymbol{\theta}_B)\|_2 \;&\leq\; \delta_B\, \tbinom{K}{2}\, \widetilde{\lambda}_{\min,B}^{\epsilon}
\end{aligned} \tag{17}$$

occur with probability at least $1 - N^{-2}$. Define, for all $t > 0$, the events

$$\begin{aligned}
\mathcal{D}_W(t) \;&:=\; \left\{\boldsymbol{x} \in \mathbb{X} : \sup_{\boldsymbol{\theta}_W \in \partial\mathcal{B}_2(\boldsymbol{\theta}_W^\star, \delta_W)} \|\nabla_{\boldsymbol{\theta}_W} m_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W) - \nabla_{\boldsymbol{\theta}_W} M_W(\boldsymbol{\theta}_W)\|_2 \geq t\right\} \\
\mathcal{D}_B(t) \;&:=\; \left\{\boldsymbol{x} \in \mathbb{X} : \sup_{\boldsymbol{\theta}_B \in \partial\mathcal{B}_2(\boldsymbol{\theta}_B^\star, \delta_B)} \|\nabla_{\boldsymbol{\theta}_B} m_B(\boldsymbol{\theta}_B, \boldsymbol{X}_B) - \nabla_{\boldsymbol{\theta}_B} M_B(\boldsymbol{\theta}_B)\|_2 \geq t\right\}.
\end{aligned}$$

By Lemma 6.2,

$$\begin{aligned}
\sup_{\boldsymbol{\theta}_W \in \partial\mathcal{B}_2(\boldsymbol{\theta}_W^\star, \delta_W)} \|\nabla_{\boldsymbol{\theta}_W} m_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W) - \nabla_{\boldsymbol{\theta}_W} M_W(\boldsymbol{\theta}_W)\|_2 \;&=\; \|\nabla_{\boldsymbol{\theta}_W} \ell_W(\boldsymbol{\theta}_W^\star, \boldsymbol{X}_W)\|_2 \\
\sup_{\boldsymbol{\theta}_B \in \partial\mathcal{B}_2(\boldsymbol{\theta}_B^\star, \delta_B)} \|\nabla_{\boldsymbol{\theta}_B} m_B(\boldsymbol{\theta}_B, \boldsymbol{X}_B) - \nabla_{\boldsymbol{\theta}_B} M_B(\boldsymbol{\theta}_B)\|_2 \;&=\; \|\nabla_{\boldsymbol{\theta}_B} \ell_B(\boldsymbol{\theta}_B^\star, \boldsymbol{X}_B)\|_2,
\end{aligned}$$

and applying Lemma 1.1, we obtain the bounds

$$\mathbb{P}\left(\mathcal{D}_W\left(\delta_W\, K\, \widetilde{\lambda}_{\min,W}^{\epsilon}\right)\right) \;\leq\; \exp\left(-\frac{\delta_W^2\, (\widetilde{\lambda}_{\min,W}^{\epsilon})^2\, K^2}{5\, K\, \widetilde{\lambda}_{\max,W}^{\star} + C_W\, A_{\max}^2\, \sqrt{p}\, \delta_W\, \widetilde{\lambda}_{\min,W}^{\epsilon}\, K} + \log(5)\, p\right)$$

and

$$\mathbb{P}\left(\mathcal{D}_B\left(\delta_B\, \tbinom{K}{2}\, \widetilde{\lambda}_{\min,B}^{\epsilon}\right)\right) \;\leq\; \exp\left(-\frac{\delta_B^2\, (\widetilde{\lambda}_{\min,B}^{\epsilon})^2\, \tbinom{K}{2}^2}{5\, \tbinom{K}{2}\, \widetilde{\lambda}_{\max,B}^{\star} + 2\, C_B\, A_{\max}^2\, \sqrt{q}\, \delta_B\, \widetilde{\lambda}_{\min,B}^{\epsilon}\, \tbinom{K}{2}} + \log(5)\, q\right).$$

Choosing

$$\delta_W \;=\; \beta_W \, \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{\frac{p}{K}} \;>\; 0,$$

for a value of $\beta_W \in (0, \infty)$ to be given, establishes

$$\mathbb{P}\left(\mathcal{D}_W \left(\delta_W \, \widetilde{\lambda}^{\epsilon}_{\min,W} \, K\right)\right) \;\leq\; \exp\left(-\frac{\beta_W^2 \; p \, K \, \widetilde{\lambda}^{\star}_{\max,W}}{5 \, K \, \widetilde{\lambda}^{\star}_{\max,W} + C_W \, \beta_W \, A_{\max}^2 \, p \sqrt{K \, \widetilde{\lambda}^{\star}_{\max,W}}} + \log(5)\, p\right).$$

Using Assumption 4, the assumption that

$$A_{\max} \;\leq\; \left(\frac{N \, \widetilde{\lambda}^{\star}_{\max,W}}{A_{\text{avg}} \, p^2}\right)^{1/4} \quad \text{implies} \quad A_{\max}^2 \, p \sqrt{K \, \widetilde{\lambda}^{\star}_{\max,W}} \;\leq\; K \, \widetilde{\lambda}^{\star}_{\max,W},$$

defining $A_{\text{avg}} \coloneqq K^{-1} \sum_{k=1}^{K} |\mathcal{A}_k|$ and using the identity

$$N \;=\; \sum_{k=1}^{K} |\mathcal{A}_k| \;=\; K \, \frac{1}{K} \sum_{k=1}^{K} |\mathcal{A}_k| \;=\; K \, A_{\text{avg}}. \tag{18}$$

Hence,

$$\begin{aligned}
&\exp\left(-\frac{\beta_W^2 \; p \, K \, \widetilde{\lambda}^{\star}_{\max,W}}{5 \, K \, \widetilde{\lambda}^{\star}_{\max,W} + C_W \, \beta_W \, A_{\max}^2 \, p \sqrt{K \, \widetilde{\lambda}^{\star}_{\max,W}}} + \log(5)\, p\right) \\
\leq\;\; &\exp\left(-\frac{\beta_W^2 \; p \, K \, \widetilde{\lambda}^{\star}_{\max,W}}{(5 + \beta_W \, C_W) \, K \, \widetilde{\lambda}^{\star}_{\max,W}} + \log(5)\, p\right) \\
=\;\; &\exp\left(\left(-\frac{\beta_W^2}{5 + \beta_W \, C_W} + \log(5)\right) p\right).
\end{aligned}$$

To obtain the desired probability guarantee, we require

$$-\frac{\beta_W^2}{5 + \beta_W \, C_W} + \log(5) \;=\; -2,$$

which in turn requires a solution $\beta_W \in (0, \infty)$ to the quadratic equation

$$\beta_W^2 - C_W \, (2 + \log(5)) \, \beta_W - 5 \, (2 + \log(5)) \;=\; 0.$$

Using the quadratic formula, such a root is given by

$$\beta_W \;=\; \frac{C_W \, (2 + \log(5)) + \sqrt{C_W^2 \, (2 + \log(5))^2 + 20 \, (2 + \log(5))}}{2} \;>\; 0,$$

which in turn establishes that

$$\mathbb{P}\left(\mathcal{D}_W\left(\delta_W\,\widetilde{\lambda}^{\epsilon}_{\min,W}\,K\right)\right) \;\;\leq\;\; \exp\left(-2\,p\right).$$

Under the assumption that $p \geq \log(N)$,

$$\mathbb{P}\left(\mathcal{D}_W\left(\delta_W\,\widetilde{\lambda}^{\epsilon}_{\min,W}\,K\right)\right) \;\;\leq\;\; \frac{1}{N^2}.$$

Similarly, choosing

$$\delta_B \;\;=\;\; \beta_B\,\frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}}\,\sqrt{\frac{q}{\binom{K}{2}}} \;\;>\;\; 0,$$

for a value of $\beta_B \in (0,\infty)$ to be given, establishes

$$\mathbb{P}\left(\mathcal{D}_B\left(\delta_B\,\widetilde{\lambda}^{\epsilon}_{\min,B}\,\tbinom{K}{2}\right)\right) \;\;\leq\;\; \exp\left(-\frac{\beta_B^2\,q\,\binom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B}}{5\,\binom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B} + 2\,C_B\,\beta_B\,A^2_{\max}\,q\,\sqrt{\binom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B}}} + \log(5)\,q\right).$$

Using Assumption 4, the assumption that

$$A_{\max} \;\leq\; \left(\frac{N^2\,\widetilde{\lambda}^{\star}_{\max,B}}{4\,A^2_{\text{avg}}\,q^2}\right)^{1/4} \qquad \text{implies} \qquad A^2_{\max}\,q\,\sqrt{\tbinom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B}} \;\leq\; \tbinom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B},$$

once more using the identity in (18). Hence,

$$\begin{aligned}
&\exp\left(-\frac{\beta_B^2\,q\,\binom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B}}{5\,\binom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B} + 2\,C_B\,\beta_B\,A^2_{\max}\,q\,\sqrt{\binom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B}}} + \log(5)\,q\right)\\
\leq\;\;&\exp\left(-\frac{\beta_B^2\,q\,\binom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B}}{(5+2\,C_B\,\beta_B)\,\binom{K}{2}\,\widetilde{\lambda}^{\star}_{\max,B}} + \log(5)\,q\right)\\
=\;\;&\exp\left(\left(-\frac{\beta_B^2}{5+2\,C_B\,\beta_B} + \log(5)\right)q\right).
\end{aligned}$$

To obtain the desired probability guarantee, we require

$$-\frac{\beta_B^2}{5+2\,C_B\,\beta_B} + \log(5) \;\;=\;\; -2,$$

which in turn requires a solution $\beta_B \in (0,\infty)$ to the quadratic equation

$$\beta_B^2 - 2\,C_B\,(2+\log(5))\,\beta_B - 5\,(2+\log(5)) \;\;=\;\; 0.$$

Using the quadratic formula, such a root is given by

$$\beta_B \;\;=\;\; C_B\,(2+\log(5)) + \sqrt{C_B^2\,(2+\log(5))^2 + 5\,(2+\log(5))} \;\;>\;\; 0,$$

which in turn establishes that

$$\mathbb{P}\left(\mathcal{D}_B\left(\delta_W\, \widetilde{\lambda}^{\epsilon}_{\min,B}\binom{K}{2}\right)\right) \;\leq\; \exp(-2\,q)\,.$$

Under the assumption that $q \geq \log(N)$,

$$\mathbb{P}\left(\mathcal{D}_W\left(\delta_B\, \widetilde{\lambda}^{\epsilon}_{\min,B}\binom{K}{2}\right)\right) \;\leq\; N^{-2}.$$

As a result, event (17) occurs with probability at least $1 - N^{-2}$, implying that, with probability at least $1 - N^{-2}$, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_W, \widehat{\boldsymbol{\theta}}_B)$ exists uniquely and satisfies

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}^\star_W\|_2 \;&\leq\; \beta_W\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{\frac{p}{K}} \\
\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}^\star_B\|_2 \;&\leq\; \beta_B\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}} \sqrt{\frac{q}{\binom{K}{2}}}.
\end{aligned}$$

Uniqueness of $(\widehat{\boldsymbol{\theta}}_W, \widehat{\boldsymbol{\theta}}_B)$ follows from the assumption that the exponential-family local dependence random graph model is minimal (Proposition 3.13 of Sundberg (2019)). We convert the bounds in terms of K to N by utilizing (18) again to show that

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}^\star_W\|_2 \;&\leq\; \beta_W\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{\frac{p}{K}} \;=\; C_1\, \sqrt{A_{\mathrm{avg}}}\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{\frac{p}{N}} \\
\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}^\star_B\|_2 \;&\leq\; \beta_B\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}} \sqrt{\frac{q}{\binom{K}{2}}} \;\leq\; C_2\, A_{\mathrm{avg}}\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}} \sqrt{\frac{q}{N^2}},
\end{aligned}$$

using $\binom{K}{2} \geq K^2/4$ in the second case, and defining $C_1 \coloneqq \beta_W > 0$ and $C_2 \coloneqq 2\,\beta_B > 0$. Both are independent of N, p, and q.

Finally, we show the restriction to $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ to be legitimate. Assumption 3 ensures

$$\lim_{N\to\infty} \sqrt{A_{\mathrm{avg}}}\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{\frac{p}{N}} \;=\; 0 \quad \text{and} \quad \lim_{N\to\infty} A_{\mathrm{avg}}\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}} \sqrt{\frac{q}{N^2}} \;=\; 0.$$

As such, there exists an $N_0 \geq 3$ such that, for all integers $N \geq N_0$, we have $\max\{\delta_W, \delta_B\} < \epsilon/\sqrt{2}$. Thus, for all integers $N \geq N_0$ and with probability at least $1 - N^{-2}$, the unique vector $\widehat{\boldsymbol{\theta}} \in \widehat{\boldsymbol{\Theta}}$ satisfies

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;=\; \sqrt{\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}^\star_W\|_2^2 + \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}^\star_B\|_2^2} \;\leq\; \sqrt{\delta_W^2 + \delta_B^2} \;<\; \epsilon,$$

which implies, for all integers $N \geq N_0$, that

$$\mathbb{P}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \leq \epsilon) \;\geq\; 1 - N^{-2},$$

justifying the restriction to the subset of the parameter space $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon) \subset \mathbb{R}^{p+q}$. $\square$

1.1. Auxiliary results for Theorem 2.1

We prove a concentration inequality for gradients of the log-likelihood which is utilized in the proof of Theorem 2.1.

Lemma 1.1. *Under the assumptions of Theorem 2.1,*

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\theta}_W}\, \ell_W(\boldsymbol{\theta}_W^\star, \boldsymbol{X}_W)\|_2 \,\leq\, \delta\right) \;\geq\; 1 - \exp\left(-\frac{\delta^2}{5\,K\,\widetilde{\lambda}^\star_{\max,W} + C_W\,A_{\max}^2\,\sqrt{p}\,\delta} + \log(5)\,p\right)$$

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\theta}_B}\, \ell_B(\boldsymbol{\theta}_B^\star, \boldsymbol{X}_B)\|_2 \,\leq\, \delta\right) \;\geq\; 1 - \exp\left(-\frac{\delta^2}{5\,\binom{K}{2}\,\widetilde{\lambda}^\star_{\max,B} + 2\,C_B A_{\max}^2\,\sqrt{q}\,\delta} + \log(5)\,q\right),$$

for all $\delta > 0$, where $C_W > 0$ and $C_B > 0$ are the same constants as in Assumption 1.

PROOF OF LEMMA 1.1. We first prove the result for the within-block case, and then discuss extensions to prove the result for the between-block case, noting that the two proofs are essentially the same with only a couple of notational changes.

Following the method utilized in the proof of Lemma 8.4 of Chen, Gao and Zhang (2022), define $\mathscr{U} := \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_2 \leq 1\}$ to be the closed unit ball in $\mathbb{R}^p$. By Corollary 4.2.13 (p. 78) of Vershynin (2018), there exists a subset $\mathscr{V}_\epsilon \subset \mathscr{U}$ (for $\epsilon \in (0,1)$) which is an ϵ-net of $\mathscr{U} \subset \mathbb{R}^p$ such that the cardinality of the set $\mathscr{V}_\epsilon$ satisfies $\log|\mathscr{V}_\epsilon| \leq p\,\log(2\,\epsilon^{-1}+1)$. Taking $\epsilon = 1/2$, there exists, for each $\boldsymbol{u} \in \mathscr{U}$, a $\boldsymbol{v} \in \mathscr{V}_{1/2}$ satisfying $\|\boldsymbol{u}-\boldsymbol{v}\|_2 \leq 1/2$, where $\log|\mathscr{V}_{1/2}| \leq p\,\log(5)$. For ease of presentation, define

$$\boldsymbol{G} \;:=\; \nabla_{\boldsymbol{\theta}_W}\, \ell_W(\boldsymbol{\theta}_W^\star, \boldsymbol{X}_W).$$

For any $\boldsymbol{u} \in \mathscr{U}$, with the corresponding $\boldsymbol{v} \in \mathscr{V}_{1/2}$, the Cauchy-Schwarz inequality implies

$$\begin{aligned} \langle \boldsymbol{u}, \boldsymbol{G}\rangle &= \langle \boldsymbol{v}, \boldsymbol{G}\rangle + \langle \boldsymbol{u}-\boldsymbol{v}, \boldsymbol{G}\rangle \\ &\leq \langle \boldsymbol{v}, \boldsymbol{G}\rangle + \|\boldsymbol{u}-\boldsymbol{v}\|_2\,\|\boldsymbol{G}\|_2 \\ &\leq \langle \boldsymbol{v}, \boldsymbol{G}\rangle + \frac{1}{2}\,\|\boldsymbol{G}\|_2, \end{aligned} \tag{19}$$

using the fact that $\|\boldsymbol{u}-\boldsymbol{v}\|_2 \leq 1/2$ in the last line. Next, choosing

$$u_i \;=\; \frac{G_i}{\|\boldsymbol{G}\|_2}, \qquad i \in \{1,\ldots,p\},$$

ensures that $\|\boldsymbol{u}\|_2 \leq 1$ so that the chosen $\boldsymbol{u}$ exists in $\mathscr{U}$. By writing

$$\langle \boldsymbol{u}, \boldsymbol{G}\rangle \;=\; \frac{1}{\|\boldsymbol{G}\|_2}\sum_{i=1}^{p} G_i^2 \;=\; \frac{\|\boldsymbol{G}\|_2^2}{\|\boldsymbol{G}\|_2} \;=\; \|\boldsymbol{G}\|_2,$$

we revisit (19) to obtain, using the above identity and re-arrangement, the inequality

$$\|\boldsymbol{G}\|_2 \;\leq\; 2\,\max_{\boldsymbol{v}\in\mathscr{V}_{1/2}}\,\langle \boldsymbol{v}, \boldsymbol{G}\rangle, \tag{20}$$

where we take the maximum over $\boldsymbol{v} \in \mathscr{V}_{1/2}$ since we cannot be sure which $\boldsymbol{v} \in \mathscr{V}_{1/2}$ would correspond to our choice of $\boldsymbol{u} \in \mathscr{U}$ above. A quick remark is in order regarding the case when $\|\boldsymbol{G}\|_2 = 0$. Note that the above implicitly assumed $\|\boldsymbol{G}\|_2 \neq 0$. In the event where $\|\boldsymbol{G}\|_2 = 0$, the inequality (20) remains true trivially, because $\boldsymbol{G} = \boldsymbol{0}_p$ and $\langle \boldsymbol{v}, \boldsymbol{G} \rangle = 0$ for all $\boldsymbol{v} \in \mathscr{V}_{1/2}$. As a result of (20) and for $\delta > 0$,

$$\begin{aligned} \mathbb{P}\left(\|\nabla_{\boldsymbol{\theta}_W} \ell_W(\boldsymbol{\theta}^\star, \boldsymbol{X})\|_2 \leq \delta\right) &\geq \mathbb{P}\left(2 \max_{\boldsymbol{v} \in \mathscr{V}_{1/2}} \langle \boldsymbol{v}, \boldsymbol{G} \rangle \leq \delta\right) \\ &= 1 - \mathbb{P}\left(2 \max_{\boldsymbol{v} \in \mathscr{V}_{1/2}} \langle \boldsymbol{v}, \boldsymbol{G} \rangle > \delta\right). \end{aligned}$$

We next focus on bounding the probability

$$\mathbb{P}\left(2 \max_{\boldsymbol{v} \in \mathscr{V}_{1/2}} \langle \boldsymbol{v}, \boldsymbol{G} \rangle > \delta\right) \leq \exp(p \log(5)) \max_{\boldsymbol{v} \in \mathscr{V}_{1/2}} \mathbb{P}\left(\langle \boldsymbol{v}, \boldsymbol{G} \rangle > \frac{\delta}{2}\right),$$

where the inequality follows from a union bound over the set of $\boldsymbol{v} \in \mathscr{V}_{1/2}$ and using the fact that $\log |\mathscr{V}_{1/2}| \leq p \log(5)$. For a given $\boldsymbol{v} \in \mathscr{V}_{1/2}$, Lemmas 6.1 and 6.2 allow us to write

$$\langle \boldsymbol{v}, \boldsymbol{G} \rangle = \sum_{i=1}^{p} v_i \left[\nabla_{\boldsymbol{\theta}_W} \ell_W(\boldsymbol{\theta}_W^\star, \boldsymbol{X}_W)\right]_i = \sum_{i=1}^{p} v_i \sum_{k=1}^{K} \left[s_{k,k,i}(\boldsymbol{X}_{k,k}) - \mathbb{E}\, s_{k,k,i}(\boldsymbol{X}_{k,k})\right].$$

Observe the following two key facts:

1. (Mean zero) The sum of random variables $\langle \boldsymbol{v}, \boldsymbol{G} \rangle$ satisfies $\mathbb{E}\, \langle \boldsymbol{v}, \boldsymbol{G} \rangle = 0$.
2. (Sum of independent random variables) The sum of random variables

$$\langle \boldsymbol{v}, \boldsymbol{G} \rangle = \sum_{k=1}^{K} \left[\sum_{i=1}^{p} v_i \left[s_{k,k,i}(\boldsymbol{X}_{k,k}) - \mathbb{E}\, s_{k,k,i}(\boldsymbol{X}_{k,k})\right]\right]$$

is a sum of mean zero independent random variables for fixed $\boldsymbol{v} \in \mathscr{V}_{1/2}$, because, by the local dependence assumption, the collection of random variables

$$\sum_{i=1}^{p} v_i \left[s_{k,k,i}(\boldsymbol{X}_{k,k}) - \mathbb{E}\, s_{k,k,i}(\boldsymbol{X}_{k,k})\right], \qquad k \in \{1, \ldots, K\}, \tag{21}$$

is a collection of independent random variables for fixed $\boldsymbol{v} \in \mathscr{V}_{1/2}$.

Together, these two points ensure the assumptions of Bernstein's inequality are met (e.g., Theorem 2.8.4, p. 35, Vershynin, 2018). Along this path, we first evaluate the variance term by writing

$$\begin{aligned}
\mathbb{V}\langle \boldsymbol{v}, \boldsymbol{G}\rangle &= \sum_{k=1}^{K} \mathbb{V}\left(\sum_{i=1}^{p} v_i \left[s_{k,k,i}(\boldsymbol{X}_{k,k}) - \mathbb{E}\, s_{k,k,i}(\boldsymbol{X}_{k,k})\right]\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathbb{C}(v_i\, s_{k,k,i}(\boldsymbol{X}_{k,k}),\, v_j\, s_{k,k,j}(\boldsymbol{X}_{k,k})) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{j=1}^{p} v_i\, v_j\, \mathbb{C}(s_{k,k,i}(\boldsymbol{X}_{k,k}),\, s_{k,k,j}(\boldsymbol{X}_{k,k})) \\
&= \sum_{k=1}^{K} \left\langle \boldsymbol{v},\, I_{k,k}(\boldsymbol{\theta}^\star)\, \boldsymbol{v}\right\rangle \\
&= \left\langle \boldsymbol{v},\, I_W(\boldsymbol{\theta}^\star)\, \boldsymbol{v}\right\rangle \\
&\leq \lambda_{\max}(I_W(\boldsymbol{\theta}^\star))\, \|\boldsymbol{v}\|_2^2 \\
&\leq \frac{9}{4}\, \lambda_{\max}(I_W(\boldsymbol{\theta}^\star)),
\end{aligned}$$

where $\lambda_{\max}(I_W(\boldsymbol{\theta}^\star))$ is the largest eigenvalue of $I_W(\boldsymbol{\theta}^\star)$, and using the inequality

$$\|\boldsymbol{v}\|_2 \;\leq\; \|\boldsymbol{u}\|_2 + \|\boldsymbol{u}-\boldsymbol{v}\|_2 \;\leq\; 1 + \frac{1}{2} \;\leq\; \frac{3}{2},$$

where the construction of the ϵ-net $\mathscr{V}_{1/2}$ of $\mathscr{U}$ with $\epsilon = 1/2$ ensures the existence of such a $\boldsymbol{u} \in \mathscr{U}$ to make the above inequality valid. This yields the final inequality

$$\max_{\boldsymbol{v}\in\mathscr{V}_{1/2}} \mathbb{V}\langle \boldsymbol{v}, \boldsymbol{G}\rangle \;\leq\; \frac{9}{4}\, \lambda_{\max}(I_W(\boldsymbol{\theta}^\star)) \;=\; \frac{9}{4}\, K\, \widetilde{\lambda}^\star_{\max,W},$$

defining

$$\widetilde{\lambda}^\star_{\max,W} \;\coloneqq\; \frac{\lambda_{\max}(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\, \ell(\boldsymbol{\theta}^\star, \boldsymbol{X}))}{K} \;=\; \frac{\lambda_{\max}(I_W(\boldsymbol{\theta}^\star))}{K}.$$

We next bound the absolute value of each random variable in (21) $\mathbb{P}$-almost surely. By Assumption 1, there exists a constant $C_W > 0$, independent of N, p, and q, such that

$$\max_{\boldsymbol{x}_{k,k}\in\mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k})\|_\infty \;\leq\; C_W \binom{|\mathcal{A}_k|}{2}, \qquad k \in \{1,\ldots,K\}. \tag{22}$$

Hence, by the Cauchy-Schwarz inequality and using (22),

$$\begin{aligned}
& \max_{k \in \{1,\dots,K\}} \sup_{\boldsymbol{x}_{k,k} \in \mathbb{X}_{k,k}} \left| \sum_{i=1}^{p} v_i \, \left(s_{k,k,i}(\boldsymbol{x}_{k,k}) - \mathbb{E}\, s_{k,k,i}(\boldsymbol{X}_{k,k})\right) \right| \\
\leq \; & \max_{k \in \{1,\dots,K\}} \sup_{\boldsymbol{x}_{k,k} \in \mathbb{X}_{k,k}} \|\boldsymbol{v}\|_2 \, \|s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}\, s_{k,k}(\boldsymbol{X}_{k,k})\|_2 \\
\leq \; & \frac{3}{2} \left(\max_{k \in \{1,\dots,K\}} \sup_{\boldsymbol{x}_{k,k} \in \mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}\, s_{k,k}(\boldsymbol{X}_{k,k})\|_2 \right) \\
\leq \; & \frac{3\sqrt{p}}{2} \left(\max_{k \in \{1,\dots,K\}} \sup_{\boldsymbol{x}_{k,k} \in \mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}\, s_{k,k}(\boldsymbol{X}_{k,k})\|_\infty \right) \\
\leq \; & 3\sqrt{p} \left(\max_{k \in \{1,\dots,K\}} \sup_{\boldsymbol{x}_{k,k} \in \mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k})\|_\infty \right) \\
\leq \; & 3\sqrt{p}\, C_W \binom{A_{\max}}{2} \\
\leq \; & \frac{3\, C_W}{2}\, A_{\max}^2 \sqrt{p},
\end{aligned}$$

noting the bound $\|\boldsymbol{v}\|_2 \leq 3/2$ demonstrated above. With these bounds, we apply Bernstein's inequality (for just the upper-tail) (e.g., Theorem 2.8.4, p. 35, Vershynin, 2018) to obtain

$$\begin{aligned}
\mathbb{P}\left(\langle \boldsymbol{v}, \boldsymbol{G} \rangle > \delta\right) \;&\leq\; \exp\left(-\frac{\delta^2 / 2}{(9/4)\, K\, \widetilde{\lambda}^{\star}_{\max,W} + (3\, C_W / 2)\, A_{\max}^2 \sqrt{p}\, \delta / 3}\right) \\
&\leq\; \exp\left(-\frac{\delta^2}{5\, K\, \widetilde{\lambda}^{\star}_{\max,W} + C_W\, A_{\max}^2 \sqrt{p}\, \delta}\right).
\end{aligned}$$

Collecting results, we have shown, for $\delta > 0$, that

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\theta}_W}\, \ell_W(\boldsymbol{\theta}^{\star}_W, \boldsymbol{X}_W)\|_2 \leq \delta\right) \;\geq\; 1 - \exp\left(-\frac{\delta^2}{5\, K\, \widetilde{\lambda}^{\star}_{\max,W} + C_W\, A_{\max}^2 \sqrt{p}\, \delta} + \log(5)\, p\right).$$

Changes for the between-bock case. By a similar argument,

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\theta}_B}\, \ell_B(\boldsymbol{\theta}^{\star}_B, \boldsymbol{X}_B)\|_2 \leq \delta\right) \;\geq\; 1 - \exp\left(-\frac{\delta^2}{5\binom{K}{2}\, \widetilde{\lambda}^{\star}_{\max,B} + 2\, C_B A_{\max}^2 \sqrt{q}\, \delta} + \log(5)\, q\right),$$

for $\delta > 0$. We highlight the main changes to the above argument. First, we now take

$$\boldsymbol{G} \;\coloneqq\; \nabla_{\boldsymbol{\theta}_B}\, \ell_B(\boldsymbol{\theta}^{\star}_B, \boldsymbol{X}_B).$$

Second, the dimension p of $\boldsymbol{\theta}_W^\star \in \mathbb{R}^p$ is replaced by the dimension q of $\boldsymbol{\theta}_B^\star \in \mathbb{R}^q$. This implies that $\log|\mathscr{V}_{1/2}| \le q\log(5)$. Third,

$$\langle \boldsymbol{v}, \boldsymbol{G}\rangle \;=\; \sum_{1\le k<l\le K}\left[\sum_{i=1}^{q} v_i\,\left[s_{k,l,i}(\boldsymbol{X}_{k,l}) - \mathbb{E}\,s_{k,l,i}(\boldsymbol{X}_{k,l})\right]\right],$$

which implies

$$\mathbb{V}\,\langle \boldsymbol{v}, \boldsymbol{G}\rangle \;\le\; \frac{9}{4}\,\lambda_{\max}(I_B(\boldsymbol{\theta}^\star)) \;=\; \frac{9}{4}\binom{K}{2}\widetilde{\lambda}^\star_{\max,B},$$

defining

$$\widetilde{\lambda}^\star_{\max,B} \;:=\; \frac{\lambda_{\max}(-\mathbb{E}\,\nabla_B^2\,\ell(\boldsymbol{\theta}^\star,\boldsymbol{X}))}{\binom{K}{2}} \;=\; \frac{\lambda_{\max}(I_B(\boldsymbol{\theta}^\star))}{\binom{K}{2}}.$$

Fourth, and finally, we have the bound

$$\max_{\{k,l\}\subseteq\{1,\dots,K\}}\;\sup_{\boldsymbol{x}_{k,l}\in\mathbb{X}_{k,l}}\left|\sum_{i=1}^{q} v_i\,\left[s_{k,l,i}(\boldsymbol{x}_{k,l}) - \mathbb{E}\,s_{k,l,i}(\boldsymbol{X}_{k,l})\right]\right| \;\le\; 3\,C_B\,A_{\max}^2\,\sqrt{q}.$$

Together, these changes will yield the inequality

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\theta}_B}\,\ell_B(\boldsymbol{\theta}_B^\star,\boldsymbol{X}_B)\|_2 \le \delta\right) \;\ge\; 1-\exp\left(-\frac{\delta^2}{5\binom{K}{2}\widetilde{\lambda}^\star_{\max,B} + 2\,C_B A_{\max}^2\,\sqrt{q}\,\delta} + \log(5)\,q\right),$$

for $\delta > 0$. □

2. Proof of Theorem 2.2

Our method of proof utilizes Fano's method for lower bounding the minimax risk. We present a general argument for lower bounding the minimax risk for exponential families utilizing Fano's method and then apply the obtained general argument to our specific cases.

General argument. Consider an exponential family of densities $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^m\}$ for a finite support $\mathbb{X}$ given by

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) \;=\; h(\boldsymbol{x})\,\exp\left(\langle\boldsymbol{\theta}, s(\boldsymbol{x})\rangle - \psi(\boldsymbol{\theta})\right) \;>\; 0, \qquad \boldsymbol{x}\in\mathbb{X},$$

data-generating parameter vector $\boldsymbol{\theta}^\star$, and define the minimax risk in the ℓ_2-norm to be

$$\mathscr{R} \;:=\; \inf_{\widehat{\boldsymbol{\theta}}}\;\sup_{\boldsymbol{\theta}\in\mathbb{R}^m}\;\mathbb{E}_{\boldsymbol{\theta}}\,\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2.$$

In the case that $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0$ for some $\boldsymbol{x}\in\mathbb{X}$, we would simply reduce the support to

$$\mathbb{X}_0 \;:=\; \{\boldsymbol{x}\in\mathbb{X} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) > 0\},$$

obtaining a family of strictly positive densities on $\mathbb{X}_0$. For ease of presentation, we therefore proceed without loss of generality assuming that $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta}\in\mathbb{R}^m\}$ are strictly positive on the support $\mathbb{X}$.

Let $\epsilon > 0$ be fixed and consider $\gamma \in (0, \epsilon)$. Assume that $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \subset \mathcal{B}_2(\boldsymbol{\theta}^\star, \gamma)$ ($M \geq 2$) is a $2\,\delta$-separated set in the metric $d(\boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{v} - \boldsymbol{w}\|_2$, i.e., $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \geq 2\,\delta$ for all pairs $\{i, j\} \subseteq \{1, \ldots, M\}$. Then, by Proposition 15.12 (p. 502) of Wainwright (2019) and the discussions following, the minimax risk $\mathscr{R}$ has the lower bound

$$\mathscr{R} \;\;\geq\;\; \delta \left[1 - \frac{\mathscr{I} + \log(2)}{\log\, M}\right],$$

where

$$\mathscr{I} \;\;\coloneqq\;\; \max_{\{i,j\} \subseteq \{1,\ldots,M\}} \mathrm{KL}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j), \tag{23}$$

defining the Kullback–Leibler divergences

$$\mathrm{KL}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \;\;\coloneqq\;\; \sum_{\boldsymbol{x} \in \mathbb{X}} f_{\boldsymbol{\theta}_i}(\boldsymbol{x}) \,\log \frac{f_{\boldsymbol{\theta}_i}(\boldsymbol{x})}{f_{\boldsymbol{\theta}_j}(\boldsymbol{x})}, \qquad \{i, j\} \subseteq \{1, \ldots, M\}.$$

For an exponential family, we can express $\mathrm{KL}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ as

$$\begin{aligned}
\mathrm{KL}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \;&=\; \sum_{\boldsymbol{x} \in \mathbb{X}} f_{\boldsymbol{\theta}_i}(\boldsymbol{x}) \left[\log h(\boldsymbol{x}) - \log h(\boldsymbol{x}) + \langle \boldsymbol{\theta}_i - \boldsymbol{\theta}_j,\, s(\boldsymbol{x})\rangle - \psi(\boldsymbol{\theta}_i) + \psi(\boldsymbol{\theta}_j)\right] \\
&=\; \mathbb{E}_{\boldsymbol{\theta}_i} \langle \boldsymbol{\theta}_i - \boldsymbol{\theta}_j,\, s(\boldsymbol{X})\rangle - \psi(\boldsymbol{\theta}_i) + \psi(\boldsymbol{\theta}_j) \\
&=\; \langle \boldsymbol{\theta}_i - \boldsymbol{\theta}_j,\, \boldsymbol{\mu}(\boldsymbol{\theta}_i)\rangle - \psi(\boldsymbol{\theta}_i) + \psi(\boldsymbol{\theta}_j),
\end{aligned} \tag{24}$$

defining $\boldsymbol{\mu}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{X})$ to be the mean-value parameter map of the exponential family. Using Corollary 2.3 of Brown (1986), we perform the expansion

$$\begin{aligned}
\psi(\boldsymbol{\theta}_j) \;&=\; \psi(\boldsymbol{\theta}_i) + \langle \boldsymbol{\theta}_j - \boldsymbol{\theta}_i,\, \boldsymbol{\mu}(\boldsymbol{\theta}_i)\rangle + \frac{1}{2}\,\langle \boldsymbol{\theta}_j - \boldsymbol{\theta}_i,\, I(\dot{\boldsymbol{\theta}})\,(\boldsymbol{\theta}_j - \boldsymbol{\theta}_i)\rangle \\
&=\; \psi(\boldsymbol{\theta}_i) - \langle \boldsymbol{\theta}_i - \boldsymbol{\theta}_j,\, \boldsymbol{\mu}(\boldsymbol{\theta}_i)\rangle + \frac{1}{2}\,\langle \boldsymbol{\theta}_i - \boldsymbol{\theta}_j,\, I(\dot{\boldsymbol{\theta}})\,(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\rangle,
\end{aligned} \tag{25}$$

where $\dot{\boldsymbol{\theta}} = t\,\boldsymbol{\theta}_i + (1-t)\,\boldsymbol{\theta}_j$ (for some $t \in (0,1)$) and $I(\boldsymbol{\theta}) \coloneqq -\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{X})$ is the Fisher information matrix corresponding to $f_{\boldsymbol{\theta}}$. Combining (24) and (25),

$$\begin{aligned}
\mathrm{KL}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \;&=\; \frac{1}{2}\,\langle \boldsymbol{\theta}_i - \boldsymbol{\theta}_j,\, I(\dot{\boldsymbol{\theta}})\,(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\rangle \\
&\leq\; \frac{1}{2}\, n\, \widetilde{\lambda}^{\epsilon}_{\max}\, \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2^2 \\
&\leq\; \frac{1}{2}\, n\, \widetilde{\lambda}^{\epsilon}_{\max}\, \left(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^\star\|_2 + \|\boldsymbol{\theta}_j - \boldsymbol{\theta}^\star\|_2\right)^2 \\
&\leq\; 2\, n\, \epsilon^2\, \widetilde{\lambda}^{\epsilon}_{\max},
\end{aligned}$$

defining for a fixed $\epsilon > 0$ the quantity

$$\widetilde{\lambda}^{\epsilon}_{\max} \;\;\coloneqq\;\; \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \frac{\lambda_{\max}(I(\boldsymbol{\theta}))}{n},$$

noting that $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \subset \mathcal{B}_2(\boldsymbol{\theta}^\star, \gamma) \subset \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ by assumption. The size M of the largest possible $2\,\delta$-separated set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \subset \mathcal{B}_2(\boldsymbol{\theta}^\star, \gamma) \subset \mathbb{R}^m$ is the packing number of $\mathcal{B}_2(\boldsymbol{\theta}^\star, \gamma)$, which by

Lemma 4.2.8 (equivalence of covering and packing numbers) and Corollary 4.2.13 (covering numbers of the Euclidean ball) of Vershynin (2018), satisfies

$$M \;\geq\; \left(\frac{\gamma}{2\,\delta}\right)^{m},$$

taking the $2\,\delta$-separated set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \subset \mathcal{B}_2(\boldsymbol{\theta}^\star, \gamma)$ to be as large as possible and applying the results to a Euclidean ball of arbitrary radius $\gamma > 0$. As a result,

$$\log M \;\geq\; m\, \log(\gamma\,/\,2\,\delta).$$

Altogether, we have demonstrated the bound

$$\mathscr{R} \;\geq\; \delta \left[1 - \frac{2\,n\,\gamma^2\,\widetilde{\lambda}^{\epsilon}_{\max} + \log(2)}{m\,\log(\gamma\,/\,2\,\delta)}\right].$$

We desire that

$$\frac{2\,n\,\gamma^2\,\widetilde{\lambda}^{\epsilon}_{\max} + \log(2)}{m\,\log(\gamma\,/\,2\,\delta)} \;\leq\; \frac{1}{2},$$

in order to show that $\mathscr{R} \geq \delta\,/\,2$. Re-arranging this inequality, we have

$$\frac{4\,n\,\gamma^2\,\widetilde{\lambda}^{\epsilon}_{\max}}{m} \;\leq\; \frac{4\,n\,\gamma^2\,\widetilde{\lambda}^{\epsilon}_{\max}}{m} + \frac{2\,\log(2)}{m} \;\leq\; \log(\gamma\,/\,2) - \log(\delta),$$

and exponentiating we obtain

$$\exp\left(\frac{4\,n\,\gamma^2\,\widetilde{\lambda}^{\epsilon}_{\max}}{m}\right) \;\leq\; \frac{\gamma\,/\,2}{\delta} \;\leq\; \frac{\gamma}{\delta}.$$

This leads us to the following inequality

$$\delta \;\leq\; \gamma\, \exp\left(-\frac{4\,n\,\gamma^2\,\widetilde{\lambda}^{\epsilon}_{\max}}{m}\right).$$

Choosing

$$\gamma \;=\; C\,\sqrt{\frac{m}{n\,\widetilde{\lambda}^{\epsilon}_{\max}}},$$

for some $C > 0$ which is presumed to be fixed, but freely chosen, yields the bound

$$\delta \;\leq\; C\, \exp(-4\,C^2)\,\sqrt{\frac{m}{n\,\widetilde{\lambda}^{\epsilon}_{\max}}}. \tag{26}$$

As long as $m = O(n\,\widetilde{\lambda}^{\epsilon}_{\max})$, we can choose $C > 0$ to ensure that $\gamma \in (0, \epsilon)$. Thus, for all $\delta > 0$ satisfying (26), we can lower bound the minimax risk $\mathscr{R}$ by

$$\mathscr{R} \;\geq\; \frac{\delta}{2}.$$

The remainder of the proof will utilize this general argument to lower bound the minimax risk in the ℓ_2-norm for exponential-family local dependence random graphs.

Lower bounds to the minimax risk in the ℓ_2-norm for exponential-family local dependence random graphs. We will first handle the within-block case by considering

$$\mathscr{R}_{W,N} \;\; := \;\; \inf_{\widehat{\boldsymbol{\theta}}_W} \sup_{\boldsymbol{\theta} \in \mathbb{R}^{p+q}} \; \mathbb{E}_{\boldsymbol{\theta}} \, \| \widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W \|_2.$$

Fix $\epsilon > 0$, independent of N, p, and q, and define

$$\widetilde{\lambda}_{\max,W}^{\epsilon} \;\; := \;\; \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^{\star}, \epsilon)} \; \frac{\lambda_{\max}\left(-\mathbb{E}\, \nabla_{\boldsymbol{\theta}_W}^2 \; \ell(\boldsymbol{\theta}, \boldsymbol{X})\right)}{K}.$$

With this definition, we revisit (26) taking $m = p$ and $n = K$ to obtain

$$\mathscr{R}_{W,N} \;\; \geq \;\; C_1 \; \exp(-4\, C_1^2) \, \sqrt{\frac{p}{K\, \widetilde{\lambda}_{\max,W}^{\epsilon}}},$$

for some $C_1 > 0$ assumed to be fixed, but freely chosen. Using the relationship

$$N \;\; = \;\; \sum_{k=1}^{K} |\mathcal{A}_k| \;\; = \;\; K \, \frac{1}{K} \sum_{k=1}^{K} |\mathcal{A}_k| \;\; = \;\; K\, A_{\text{avg}},$$

defining $A_{\text{avg}} := K^{-1} \sum_{k=1}^{K} |\mathcal{A}_k|$, we obtain

$$\mathscr{R}_{W,N} \;\; \geq \;\; C_1 \; \exp(-4\, C_1^2) \, \sqrt{\frac{A_{\text{avg}}}{\widetilde{\lambda}_{\max,W}^{\epsilon}}} \, \sqrt{\frac{p}{N}}.$$

Then there exists $B_1 := C_1 \exp(-4\, C_1^2) > 0$, independent of N, p, and q, such that

$$\begin{aligned}
\mathscr{R}_{W,N} \;\; &\geq \;\; B_1 \sqrt{\frac{A_{\text{avg}}}{\widetilde{\lambda}_{\max,W}^{\epsilon}}} \, \sqrt{\frac{p}{N}} \\
&= \;\; B_1 \, \frac{1}{\sqrt{\widetilde{\lambda}_{\max,W}^{\epsilon}}} \, \sqrt{\frac{\widetilde{\lambda}_{\max,W}^{\star}}{\widetilde{\lambda}_{\max,W}^{\star}}} \, \frac{\widetilde{\lambda}_{\min,W}^{\epsilon}}{\widetilde{\lambda}_{\min,W}^{\epsilon}} \, \sqrt{A_{\text{avg}}} \, \sqrt{\frac{p}{N}} \\
&= \;\; B_1 \left(\frac{\widetilde{\lambda}_{\min,W}^{\epsilon}}{\sqrt{\widetilde{\lambda}_{\max,W}^{\star} \; \widetilde{\lambda}_{\max,W}^{\epsilon}}} \right) \frac{\sqrt{\widetilde{\lambda}_{\max,W}^{\star}}}{\widetilde{\lambda}_{\min,W}^{\epsilon}} \, \sqrt{A_{\text{avg}}} \, \sqrt{\frac{p}{N}} \\
&\geq \;\; B_1 \left(\frac{\widetilde{\lambda}_{\min,W}^{\epsilon}}{\widetilde{\lambda}_{\max,W}^{\epsilon}} \right) \frac{\sqrt{\widetilde{\lambda}_{\max,W}^{\star}}}{\widetilde{\lambda}_{\min,W}^{\epsilon}} \, \sqrt{A_{\text{avg}}} \, \sqrt{\frac{p}{N}}.
\end{aligned}$$

The above inequality establishes

$$\mathscr{R}_{W,N} \;\;\geq\;\; B_1 \sqrt{\frac{A_{\text{avg}}}{\widetilde{\lambda}^{\epsilon}_{\max,W}}} \sqrt{\frac{p}{N}} \;\;\geq\;\; B_1 \left(\frac{\widetilde{\lambda}^{\epsilon}_{\min,W}}{\widetilde{\lambda}^{\epsilon}_{\max,W}}\right) \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{A_{\text{avg}}} \sqrt{\frac{p}{N}}.$$

Next, we prove the between-block case and consider

$$\mathscr{R}_{B,N} \;\;\coloneqq\;\; \inf_{\widehat{\boldsymbol{\theta}}_B} \sup_{\boldsymbol{\theta} \in \mathbb{R}^{p+q}} \mathbb{E}_{\boldsymbol{\theta}} \, \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B\|_2.$$

Fix $\epsilon > 0$, independent of N, p, and q, and define

$$\widetilde{\lambda}^{\epsilon}_{\max,B} \;\;\coloneqq\;\; \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star,\epsilon)} \frac{\lambda_{\max}\left(-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_B} \ell(\boldsymbol{\theta}, \boldsymbol{X})\right)}{\binom{K}{2}}.$$

With this definition, we revisit (26) taking $m = q$ and $n = \binom{K}{2}$ to obtain

$$\mathscr{R}_{B,N} \;\;\geq\;\; C_2 \, \exp(-4\, C_2^2) \sqrt{\frac{q}{\binom{K}{2}\, \widetilde{\lambda}^{\epsilon}_{\max,B}}},$$

for some $C_2 > 0$ assumed to be fixed, but freely chosen. Using the relationship

$$N \;\;=\;\; \sum_{k=1}^{K} |\mathcal{A}_k| \;\;=\;\; K \, \frac{1}{K} \sum_{k=1}^{K} |\mathcal{A}_k| \;\;=\;\; K\, A_{\text{avg}},$$

defining $A_{\text{avg}} \coloneqq K^{-1} \sum_{k=1}^{K} |\mathcal{A}_k|$, we obtain

$$\begin{aligned}
\mathscr{R}_{B,N} \;\;&\geq\;\; C_2 \, \exp(-4\, C_2^2) \, \frac{1}{\sqrt{\widetilde{\lambda}^{\epsilon}_{\max,B}}} \sqrt{\frac{q}{\binom{N \,/\, A_{\text{avg}}}{2}}} \\
&\geq\;\; C_2 \, \exp(-4\, C_2^2) \, \frac{1}{\sqrt{\widetilde{\lambda}^{\epsilon}_{\max,B}}} \sqrt{\frac{2\, q\, A_{\text{avg}}^2}{N^2}} \;\;=\;\; B_2 \, \frac{A_{\text{avg}}}{\sqrt{\widetilde{\lambda}^{\epsilon}_{\max,B}}} \sqrt{\frac{q}{N^2}},
\end{aligned}$$

where $B_2 \coloneqq \sqrt{2}\, C_2\, \exp(-4\, C_2^2) > 0$ is independent of N, p, and q. Hence,

$$\begin{aligned}
\mathscr{R}_{B,N} &\geq B_2\, \frac{A_{\mathrm{avg}}}{\sqrt{\widetilde{\lambda}^{\epsilon}_{\max,W}}} \sqrt{\frac{q}{N^2}} \\
&= B_2\, \frac{1}{\sqrt{\widetilde{\lambda}^{\epsilon}_{\max,W}}} \sqrt{\frac{\widetilde{\lambda}^{\star}_{\max,W}}{\widetilde{\lambda}^{\star}_{\max,W}}}\, \frac{\widetilde{\lambda}^{\epsilon}_{\min,W}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\, A_{\mathrm{avg}} \sqrt{\frac{q}{N^2}} \\
&= B_2 \left(\frac{\widetilde{\lambda}^{\epsilon}_{\min,W}}{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}\, \widetilde{\lambda}^{\epsilon}_{\max,W}}} \right) \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\, A_{\mathrm{avg}} \sqrt{\frac{q}{N^2}} \\
&\geq B_2 \left(\frac{\widetilde{\lambda}^{\epsilon}_{\min,W}}{\widetilde{\lambda}^{\epsilon}_{\max,W}} \right) \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\, A_{\mathrm{avg}} \sqrt{\frac{q}{N^2}},
\end{aligned}$$

showing the claimed result of

$$\mathscr{R}_{B,N} \;\geq\; B_2\, \frac{A_{\mathrm{avg}}}{\sqrt{\widetilde{\lambda}^{\epsilon}_{\max,W}}} \sqrt{\frac{q}{N^2}} \;\geq\; B_2 \left(\frac{\widetilde{\lambda}^{\epsilon}_{\min,W}}{\widetilde{\lambda}^{\epsilon}_{\max,W}} \right) \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\, A_{\mathrm{avg}} \sqrt{\frac{q}{N^2}}.$$

Lastly, the assumption in the general argument that $m = O(n\, \widetilde{\lambda}^{\epsilon}_{\max})$ requires:

- $p = \dim(\boldsymbol{\theta}^{\star}_W)$ satisfies $p = O(N\, \widetilde{\lambda}^{\epsilon}_{\max,W})$, and
- $q = \dim(\boldsymbol{\theta}^{\star}_B)$ satisfies $q = O(N^2\, \widetilde{\lambda}^{\epsilon}_{\max,B})$,

substituting the relevant quantities into $m = O(n\, \widetilde{\lambda}^{\epsilon}_{\max})$ for each case. □

3. Proof of Theorem 2.3

We start by considering the restricted minimax risk

$$\inf_{\widehat{\boldsymbol{\theta}}_W} \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^{\star}, \epsilon)} \mathbb{E}_{\boldsymbol{\theta}}\, \| \widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W \|_2,$$

where $\epsilon > 0$ is the same as in Assumptions 2 and 3. Let $\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^{\star}, \epsilon) \subset \mathbb{R}^{p+q}$ be arbitrary. We partition the support $\mathbb{X}$ of $\boldsymbol{X}$ as follows:

$$\begin{aligned}
\mathbb{X}_1 &\coloneqq \left\{ \boldsymbol{x} \in \mathbb{X} \,:\, \| \widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}^{\star}_W \|_2 \,\leq\, C_1\, \sqrt{A_{\mathrm{avg}}}\, \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{\frac{p}{N}} \right\} \\
\mathbb{X}_2 &\coloneqq \mathbb{X} \setminus \mathbb{X}_1,
\end{aligned}$$

where the constant $C_1 > 0$ is the same as the one guaranteed in Theorem 2.1. Then,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 &= \mathbb{E}_{\boldsymbol{\theta}} \left[\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \,|\, \boldsymbol{X} \in \mathbb{X}_1 \right] \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} \in \mathbb{X}_1) \\ &+ \mathbb{E}_{\boldsymbol{\theta}} \left[\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \,|\, \boldsymbol{X} \in \mathbb{X}_2 \right] \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} \in \mathbb{X}_2), \end{aligned} \tag{27}$$

by the law of total expectation, and

$$\begin{aligned} \inf_{\widehat{\boldsymbol{\theta}}_W} \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 &\leq \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \mathbb{E}_{\boldsymbol{\theta}} \left[\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \,|\, \boldsymbol{X} \in \mathbb{X}_1 \right] \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} \in \mathbb{X}_1) \\ &+ \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \mathbb{E}_{\boldsymbol{\theta}} \left[\|\widetilde{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \,|\, \boldsymbol{X} \in \mathbb{X}_2 \right] \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} \in \mathbb{X}_2), \end{aligned}$$

where $\widehat{\boldsymbol{\theta}}_W$ in the first term in the upper bound is the maximum likelihood estimator and

$$\widetilde{\boldsymbol{\theta}}_W \;\; := \;\; \underset{\boldsymbol{\theta}_W \in \mathcal{B}_2(\mathbf{0}_p, 100)}{\arg\max} \;\; \ell_W(\boldsymbol{\theta}_W, \boldsymbol{X}_W)$$

is the maximum likelihood estimator restricted to the subset $\mathcal{B}_2(\mathbf{0}_p, 100)$. Note that Theorem 2.1 establishes that there exists $N_0 \geq 1$, independent of N, p, or q, such that $\mathbb{P}(\boldsymbol{X} \in \mathbb{X}_1) \geq 1 - N^{-2}$, in which case we have the bound $\mathbb{P}(\boldsymbol{X} \in \mathbb{X}_2) \leq N^{-2}$, for all integers $N \geq N_0$. Observe that the upper bound becomes trivial in the case when $\mathbb{P}(\boldsymbol{X} \in \mathbb{X}_2) = 0$. We then obtain the upper bound

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \mathbb{E}_{\boldsymbol{\theta}} \left[\|\widetilde{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \,|\, \boldsymbol{X} \in \mathbb{X}_2 \right] \;\; \leq \;\; \sup_{\widetilde{\boldsymbol{\theta}}_W \in \mathcal{B}_2(\mathbf{0}_p, 100)} \; \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \|\widetilde{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2.$$

Define

$$M \;\; := \;\; \sup_{\widetilde{\boldsymbol{\theta}}_W \in \mathcal{B}_2(\mathbf{0}_p, 100)} \; \sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} \|\widetilde{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \;\; \in \;\; (0, \infty).$$

Continuing from (27) leads us to the upper bound

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 &\leq \mathbb{E}_{\boldsymbol{\theta}} \left[\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \,|\, \boldsymbol{X} \in \mathbb{X}_1 \right] \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} \in \mathbb{X}_1) + M\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} \in \mathbb{X}_2) \\ &\leq \mathbb{E}_{\boldsymbol{\theta}} \left[\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \,|\, \boldsymbol{X} \in \mathbb{X}_1 \right] \left(1 - \frac{1}{N^2}\right) + M \left(\frac{1}{N^2}\right) \\ &\leq \mathbb{E}_{\boldsymbol{\theta}} \left[\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \,|\, \boldsymbol{X} \in \mathbb{X}_1 \right] + \frac{M}{N^2} \\ &\leq C_1 \sqrt{A_{\mathrm{avg}}} \, \frac{\sqrt{\widetilde{\lambda}^\star_{\max, W}}}{\widetilde{\lambda}^\epsilon_{\min, W}} \sqrt{\frac{p}{N}} + \frac{M}{N^2}. \end{aligned}$$

Note that Assumption 2 implies that

$$\lim_{N \to \infty} \sqrt{A_{\mathrm{avg}}} \, \frac{\sqrt{\widetilde{\lambda}^\star_{\max, W}}}{\widetilde{\lambda}^\epsilon_{\min, W}} \sqrt{\frac{p}{N}} \;\; = \;\; 0,$$

which further implies

$$\sqrt{A_{\text{avg}}}\,\frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\,\sqrt{p} \;=\; o\left(\sqrt{N}\right),$$

in turn ultimately implying

$$\sqrt{A_{\text{avg}}}\,\frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\,\sqrt{\frac{p}{N}} \;=\; o\left(N^2\right).$$

As a result, there exists an $N_1 \geq N_0$, independent of N, p, and q, such that

$$\sqrt{A_{\text{avg}}}\,\frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\,\sqrt{\frac{p}{N}} \;\geq\; \frac{1}{N^2}, \qquad N \geq N_1.$$

Defining $C_2 \coloneqq C_1 + M > 0$, also independent of N, p, and q, we have

$$\inf_{\widehat{\boldsymbol{\theta}}_W}\;\sup_{\boldsymbol{\theta}\in\mathcal{B}_2(\boldsymbol{\theta}^{\star},\epsilon)}\;\mathbb{E}_{\boldsymbol{\theta}}\,\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W\|_2 \;\leq\; C_2\,\sqrt{A_{\text{avg}}}\,\frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\,\sqrt{\frac{p}{N}}.$$

Repeating the same essential argument for

$$\inf_{\widehat{\boldsymbol{\theta}}_B}\;\sup_{\boldsymbol{\theta}\in\mathcal{B}_2(\boldsymbol{\theta}^{\star},\epsilon)}\;\mathbb{E}_{\boldsymbol{\theta}}\,\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B\|_2$$

we obtain, for $C_3 > 0$ defined similarly to $C_2 > 0$,

$$\inf_{\widehat{\boldsymbol{\theta}}_B}\;\sup_{\boldsymbol{\theta}\in\mathcal{B}_2(\boldsymbol{\theta}^{\star},\epsilon)}\;\mathbb{E}_{\boldsymbol{\theta}}\,\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B\|_2 \;\leq\; C_3\,A_{\text{avg}}\,\frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}}\,\sqrt{\frac{q}{N^2}}.$$

□

4. Proof of Corollary 2.4

The assumptions of both Theorems 2.1 and 2.2 are met. Theorem 2.1 supplies the following upper bounds to the ℓ_2-error of maximum likelihood estimators:

$$\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}^{\star}_W\|_2 \;\leq\; C\,\sqrt{A_{\text{avg}}}\,\frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}}\,\sqrt{\frac{p}{N}}$$

$$\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}^{\star}_B\|_2 \;\leq\; C\,A_{\text{avg}}\,\frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,B}}}{\widetilde{\lambda}^{\epsilon}_{\min,B}}\,\sqrt{\frac{q}{N^2}}.$$

Theorem 2.2 provides the lower bounds on the minimax risks $\mathscr{R}_{W,N}$ and $\mathscr{R}_{B,N}$:

$$\mathscr{R}_{W,N} \quad \geq \quad B_1 \left(\frac{\widetilde{\lambda}^{\epsilon}_{\min,W}}{\widetilde{\lambda}^{\epsilon}_{\max,W}} \right) \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} \sqrt{A_{\mathrm{avg}}} \sqrt{\frac{p}{N}}$$

$$\mathscr{R}_{B,N} \quad \geq \quad B_2 \left(\frac{\widetilde{\lambda}^{\epsilon}_{\min,W}}{\widetilde{\lambda}^{\epsilon}_{\max,W}} \right) \frac{\sqrt{\widetilde{\lambda}^{\star}_{\max,W}}}{\widetilde{\lambda}^{\epsilon}_{\min,W}} A_{\mathrm{avg}} \sqrt{\frac{q}{N^2}}.$$

Inspecting the two sets of bounds, the assumption that

$$\widetilde{\lambda}^{\epsilon}_{\max,W} = O\left(\widetilde{\lambda}^{\epsilon}_{\min,W}\right) \quad \text{and} \quad \widetilde{\lambda}^{\epsilon}_{\max,B} = O\left(\widetilde{\lambda}^{\epsilon}_{\min,B}\right)$$

ensures that the two sets of bounds match (up to an unknown, but fixed, constant). As a result, the upper bounds on the ℓ_2-error presented in Theorem 2.1 achieve (up to an unknown, but fixed, constant) the minimax rate of convergence. ☐

5. Proof of Theorem 2.5

For ease of presentation, we first present a general argument for bounding the error of the multivariate normal approximation, and then show how it can be applied to maximum likelihood estimators of exponential-families of local dependence random graph models in order to establish the desired result.

Bounding the error of the multivariate normal approximation. Consider a general estimating function $m : \mathbb{R}^d \times \mathbb{X} \mapsto \mathbb{R}$ which admits the following form:

$$m(\boldsymbol{\theta}, \boldsymbol{x}) \quad = \quad \sum_{k=1}^{K} m_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) + \sum_{1 \leq k < l \leq K} m_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}),$$

and assume $m(\boldsymbol{\theta}, \boldsymbol{x})$ is thrice continuously differentiable in the elements of $\boldsymbol{\theta} \in \mathbb{R}^{p+q}$. By assumption, the subgraphs $\boldsymbol{X}_{k,l}$ $(1 \leq k \leq l \leq K)$ are mutually independent, implying, for a fixed $\boldsymbol{\theta} \in \mathbb{R}^{p+q}$, that the collection of random variables $m_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{X}_{k,k})$ $(1 \leq k \leq K)$ and $m_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{X}_{k,l})$ $(1 \leq k < l \leq K)$ are likewise mutually independent. As such,

$$\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}, \boldsymbol{x}) \quad = \quad \sum_{k=1}^{K} \nabla_{\boldsymbol{\theta}}\, m_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) + \sum_{1 \leq k < l \leq K} \nabla_{\boldsymbol{\theta}}\, m_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}) \tag{28}$$

is a sum of mutually independent random vectors. Assume that

$$\begin{aligned} \mathbb{E}\, \nabla_{\boldsymbol{\theta}}\, m_{k,k}(\boldsymbol{\theta}^{\star}_W, \boldsymbol{X}_{k,k}) \quad &= \quad \boldsymbol{0}_{p+q}, \qquad 1 \leq k \leq K \\ \mathbb{E}\, \nabla_{\boldsymbol{\theta}}\, m_{k,l}(\boldsymbol{\theta}^{\star}_B, \boldsymbol{X}_{k,l}) \quad &= \quad \boldsymbol{0}_{p+q}, \qquad 1 \leq k < l \leq K, \end{aligned} \tag{29}$$

implying that $\mathbb{E}\, \nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}^{\star}, \boldsymbol{X}) = \boldsymbol{0}_{p+q}$.

Let $\boldsymbol{\theta} \in \mathbb{R}^{p+q}$ and $\boldsymbol{x} \in \mathbb{X}$ be fixed. By a multivariate Taylor expansion,

$$\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}, \boldsymbol{X}) \quad = \quad \nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}^{\star}, \boldsymbol{X}) + \nabla^2_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}^{\star}, \boldsymbol{X})\, (\boldsymbol{\theta} - \boldsymbol{\theta}^{\star}) + \boldsymbol{R}, \tag{30}$$

where $\boldsymbol{R} \in \mathbb{R}^{p+q}$ is a vector of remainders given in the Lagrange form, where each of the remainder terms R_i ($i \in \{1, \ldots, p+q\}$) is given by

$$\begin{aligned} R_i &= \sum_{j=1}^{p+q} \frac{1}{2} \left[\frac{\partial^2}{\partial\, \theta_j^2} \left[\nabla_{\boldsymbol{\theta}}\, m(\dot{\boldsymbol{\theta}}^{(i)}, \boldsymbol{X}) \right]_i \right] (\theta_j - \theta_j^\star)^2 \\ &+ \sum_{1 \le j < r \le p+q} \frac{1}{2} \left[\frac{\partial^2}{\partial\, \theta_j\, \partial\, \theta_r} \left[\nabla_{\boldsymbol{\theta}}\, m(\dot{\boldsymbol{\theta}}^{(i)}, \boldsymbol{X}) \right]_i \right] (\theta_j - \theta_j^\star)\, (\theta_r - \theta_r^\star), \end{aligned} \tag{31}$$

where $\dot{\boldsymbol{\theta}}^{(i)} = t_i\, \boldsymbol{\theta} + (1 - t_i)\, \boldsymbol{\theta}^\star$ (for some $t_i \in [0,1]$, $i \in \{1, \ldots, p+q\}$). Assume that $\boldsymbol{C} := \mathbb{V}\, \nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}^\star, \boldsymbol{X})$ is non-singular and that $\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}, \boldsymbol{x})$ has a root given by $\boldsymbol{\theta}_0 \in \mathbb{R}^{p+q}$. Taking $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we re-arrange (30) with the observation $\boldsymbol{X} = \boldsymbol{x}$ in order to obtain

$$\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}^\star, \boldsymbol{x}) \;=\; \nabla_{\boldsymbol{\theta}}^2\, m(\boldsymbol{\theta}^\star, \boldsymbol{x})\, (\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0) - \boldsymbol{R}. \tag{32}$$

Bear in mind, from the form of (28), that $\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}^\star, \boldsymbol{x})$ is a sum of independent random vectors. Define

$$\boldsymbol{Y}_{k,l} \;:=\; \begin{cases} \boldsymbol{C}^{-1/2}\; \nabla_{\boldsymbol{\theta}}\, m_{k,k}(\boldsymbol{\theta}_W^\star, \boldsymbol{X}_{k,k}), & \text{if } k = l \\ \boldsymbol{C}^{-1/2}\; \nabla_{\boldsymbol{\theta}}\, m_{k,l}(\boldsymbol{\theta}_B^\star, \boldsymbol{X}_{k,l}), & \text{if } k \neq l \end{cases}, \qquad 1 \le k \le l \le K,$$

and

$$\boldsymbol{S} \;:=\; \sum_{1 \le k \le l \le K} \boldsymbol{Y}_{k,l}.$$

Observe that, by (29), $\mathbb{E}\, \boldsymbol{S} = \boldsymbol{0}_{p+q}$ and that, by the definition of $\boldsymbol{C}$, $\mathbb{V}\, \boldsymbol{S} = \boldsymbol{I}_{p+q}$, where $\boldsymbol{I}_{p+q}$ is the $(p+q)$-dimensional identity matrix. Applying Lemma 5.1, for all measurable convex sets $\mathscr{C} \subset \mathbb{R}^{p+q}$,

$$\begin{aligned} |\mathbb{P}(\boldsymbol{S} \in \mathscr{C}) - \Phi_{p+q}(\boldsymbol{Z}_{p+q} \in \mathscr{C})| &\le (42\, (p+q)^{1/4} + 16) \sum_{1 \le k \le l \le K} \mathbb{E}\, \|\boldsymbol{Y}_{k,l}\|_2^3 \\ &\le 58\, (p+q)^{1/4} \sum_{1 \le k \le l \le K} \mathbb{E}\, \|\boldsymbol{Y}_{k,l}\|_2^3. \end{aligned}$$

Normality results for $\boldsymbol{S}$ can be extended to a standardization of $(\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0)$ via (32):

$$\boldsymbol{S} \;\stackrel{D}{=}\; \boldsymbol{C}^{-1/2} \left[\nabla_{\boldsymbol{\theta}}^2\, m(\boldsymbol{\theta}^\star, \boldsymbol{x})\, (\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0) - \boldsymbol{R} \right], \tag{33}$$

where $\stackrel{D}{=}$ indicates equality in distribution.

Multivariate normal approximation for maximum likelihood estimators. Define

$$\begin{aligned} m_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) &:= \ell_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}), & 1 \le k \le K \\ m_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}) &:= \ell_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}), & 1 \le k < l \le K. \end{aligned}$$

We next verify that the assumptions placed on $m(\boldsymbol{\theta}, \boldsymbol{x})$ in the general argument presented above are met in the case of maximum likelihood estimation for exponential-family local dependence random graphs.

By Lemma 6.1,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\, m_{k,k}(\boldsymbol{\theta}_W, \boldsymbol{x}_{k,k}) &= (s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,k}(\boldsymbol{X}_{k,k}),\, \mathbf{0}_q),\\ \nabla_{\boldsymbol{\theta}}\, m_{k,l}(\boldsymbol{\theta}_B, \boldsymbol{x}_{k,l}) &= (\mathbf{0}_p,\, s_{k,l}(\boldsymbol{x}_{k,l}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,l}(\boldsymbol{X}_{k,l})),\end{aligned}$$

noting that $\boldsymbol{\theta} = (\boldsymbol{\theta}_W, \boldsymbol{\theta}_B) \in \mathbb{R}^{p+q}$, implying $\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}, \boldsymbol{x}) = s(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{x})$. Observe that

$$\begin{aligned}\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\, m_{k,k}(\boldsymbol{\theta}_W^\star, \boldsymbol{X}_{k,k})\right] &= \mathbf{0}_{p+q}, \qquad 1 \le k \le K\\ \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\, m_{k,l}(\boldsymbol{\theta}_B^\star, \boldsymbol{X}_{k,l})\right] &= \mathbf{0}_{p+q}, \qquad 1 \le k < l \le K,\end{aligned}$$

implying $\mathbb{E}\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}^\star, \boldsymbol{X}) = \mathbf{0}_{p+q}$. Lemma 6.1 additionally establishes that

$$\nabla_{\boldsymbol{\theta}}^2\, m(\boldsymbol{\theta}^\star, \boldsymbol{x}) \;=\; \mathbb{V}\, s(\boldsymbol{X}) \;=\; \mathbb{V}\,(s(\boldsymbol{X}) - \mathbb{E}\, s(\boldsymbol{X})) \;=\; \mathbb{V}\,\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta}^\star, \boldsymbol{X}),$$

implying $\boldsymbol{C} = \mathbb{V}\, s(\boldsymbol{X}) = \nabla_{\boldsymbol{\theta}}^2\, m(\boldsymbol{\theta}, \boldsymbol{x}) = -\mathbb{E}\,\nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}, \boldsymbol{X})$, which is non-singular for all $\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ by Assumption 2. Restricting to $\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$, we have verified all conditions placed on $m(\boldsymbol{\theta}, \boldsymbol{x})$ in the general argument outlined above. In this case, (33) can be expressed as

$$\boldsymbol{S} \;\overset{D}{=}\; I(\boldsymbol{\theta}^\star)^{1/2}\,(\boldsymbol{\theta}^\star - \boldsymbol{\theta}) - I(\boldsymbol{\theta}^\star)^{-1/2}\,\boldsymbol{R},$$

where $I(\boldsymbol{\theta}^\star) = \mathbb{V}\, s(\boldsymbol{X})$ is the Fisher information matrix evaluated at the data-generating parameter vector $\boldsymbol{\theta}^\star \in \mathbb{R}^{p+q}$. The local dependence assumption and the partitioning of the sufficient statistics vector $s(\boldsymbol{X}) = (s_W(\boldsymbol{X}_W), s_B(\boldsymbol{X}_B))$ imply that

$$I(\boldsymbol{\theta}^\star) \;=\; \mathbb{V}\, s(\boldsymbol{X}) \;=\; \begin{pmatrix} I_W(\boldsymbol{\theta}_W^\star) & \mathbf{0}_{p,q}\\ \mathbf{0}_{q,p} & I_B(\boldsymbol{\theta}_B^\star)\end{pmatrix},$$

where $\mathbf{0}_{d,r}$ is the $(d \times r)$-dimensional matrix consisting of all zeros, with the definitions

$$\begin{aligned}I_W(\boldsymbol{\theta}_W^\star) &\coloneqq \sum_{k=1}^{K} -\mathbb{E}\,\nabla_{\boldsymbol{\theta}_W}^2\, \ell_{k,k}(\boldsymbol{\theta}_W^\star, \boldsymbol{X}_{k,k})\\ I_B(\boldsymbol{\theta}_B^\star) &\coloneqq \sum_{1 \le k < l \le K} -\mathbb{E}\,\nabla_{\boldsymbol{\theta}_B}^2\, \ell_{k,l}(\boldsymbol{\theta}_B^\star, \boldsymbol{X}_{k,l}).\end{aligned}$$

The proof is completed by establishing the following two additional results.

I. **Convergence rate of the multivariate normal approximation**

We establish the convergence rate of the multivariate normal approximation by bounding $\sum_{1 \le k \le l \le K} \mathbb{E}\,\|\boldsymbol{Y}_{k,l}\|_2^3$. In order to do so, we bound each term:

$$\begin{aligned}\|\boldsymbol{Y}_{k,k}\|_2 &= \|I_W(\boldsymbol{\theta}_W^\star)^{-1/2}\left[s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,k}(\boldsymbol{X}_{k,k})\right]\|_2\\ &\le |||I_W(\boldsymbol{\theta}_W^\star)^{-1/2}|||_2\, \|s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,k}(\boldsymbol{X}_{k,k})\|_2\\ &\le \frac{\|s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,k}(\boldsymbol{X}_{k,k})\|_2}{\sqrt{K\,\widetilde{\lambda}_{\min,W}^\star}},\end{aligned}$$

using the bound on the spectral norm $|||I_W(\boldsymbol{\theta}_W^\star)^{-1/2}|||_2$ of the matrix $I_W(\boldsymbol{\theta}_W^\star)^{-1/2}$:

$$|||I_W(\boldsymbol{\theta}_W^\star)^{-1/2}|||_2 \;\leq\; \frac{1}{\sqrt{\lambda_{\min}(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell(\boldsymbol{\theta}^\star, \boldsymbol{X}))}} \;=\; \frac{1}{\sqrt{K\,\widetilde{\lambda}^\star_{\min,W}}},$$

which follows from Assumption 2, because if $\lambda_{\min}(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell(\boldsymbol{\theta}^\star, \boldsymbol{X}))$ is the smallest eigenvalue of $I_W(\boldsymbol{\theta}_W^\star) := -\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\ell(\boldsymbol{\theta}^\star, \boldsymbol{X})$, then $1\,/\,\lambda_{\min}(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_W}\,\ell(\boldsymbol{\theta}^\star, \boldsymbol{X}))$ is the largest eigenvalue of $I_W(\boldsymbol{\theta}_W^\star)^{-1}$. By Assumption 1, there exists $C_W > 0$, independent of N, p, and q, such that

$$\sup_{\boldsymbol{x}_{k,k}\in\mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k})\|_\infty \;\leq\; C_W\binom{|\mathcal{A}_k|}{2}, \qquad 1 \leq k \leq K,$$

which in turn implies the inequality

$$\begin{aligned}\|s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,k}(\boldsymbol{X}_{k,k})\|_2 \;&\leq\; \sqrt{p}\,\|s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,k}(\boldsymbol{X}_{k,k})\|_\infty \\ &\leq\; 2\sqrt{p}\,\|s_{k,k}(\boldsymbol{x}_{k,k})\|_\infty \\ &\leq\; 2\sqrt{p}\,C_W\binom{|\mathcal{A}_k|}{2} \\ &\leq\; \sqrt{p}\,C_W\,A^2_{\max},\end{aligned}$$

using the inequality $\binom{|\mathcal{A}_k|}{2} \leq |\mathcal{A}_k|^2\,/\,2 \leq A^2_{\max}\,/\,2$. Collecting the above bounds,

$$\|\boldsymbol{Y}_{k,k}\|_2^3 \;\leq\; \left(\frac{\sqrt{p}\,C_W\,A^2_{\max}}{\sqrt{K\,\widetilde{\lambda}^\star_{\min,W}}}\right)^3 \;=\; \frac{p^{3/2}\,C_W^3\,A^6_{\max}}{K^{3/2}\,(\widetilde{\lambda}^\star_{\min,W})^{3/2}},$$

which implies

$$\sum_{k=1}^{K}\mathbb{E}\,\|\boldsymbol{Y}_{k,k}\|_2^3 \;\leq\; \frac{p^{3/2}\,C_W^3\,A^6_{\max}}{K^{1/2}\,(\widetilde{\lambda}^\star_{\min,W})^{3/2}}.$$

A similar argument will reveal, using Assumption 1 once more, that

$$\sup_{\boldsymbol{x}_{k,l}\in\mathbb{X}_{k,l}} \|s_{k,l}(\boldsymbol{x}_{k,l}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,l}(\boldsymbol{X}_{k,l})\|_\infty \;\leq\; 2\,C_B\,|\mathcal{A}_k|\,|\mathcal{A}_l| \;\leq\; 2\,C_B\,A^2_{\max},$$

for $1 \leq k < l \leq K$, which will instead yield the bound

$$\sum_{1\leq k<l\leq K}\mathbb{E}\,\|\boldsymbol{Y}_{k,l}\|_2^3 \;\leq\; \frac{2\,q^{3/2}\,C_B^3\,A^6_{\max}}{\binom{K}{2}^{1/2}\,(\widetilde{\lambda}^\star_{\min,B})^{3/2}} \;\leq\; \frac{2\sqrt{2}\,q^{3/2}\,C_B^3\,A^6_{\max}}{K\,(\widetilde{\lambda}^\star_{\min,B})^{3/2}},$$

noting there are $\binom{K}{2}$ between-block subgraphs $\boldsymbol{X}_{k,l}$ ($1 \leq k < l \leq K$), as opposed to K within-block subgraphs $\boldsymbol{X}_{k,k}$ ($1 \leq k \leq K$) and using the bound $\binom{K}{2} \geq K^2\,/\,2$. Collecting terms and using the bound $N = \sum_{k=1}^K |\mathcal{A}_k| \leq K\, A_{\max}$, we obtain the bound

$$\begin{aligned}\sum_{1 \leq k \leq l \leq K} \mathbb{E}\, \|\boldsymbol{Y}_{k,l}\|_2^3 &\leq A_{\max}^6 \left[\frac{p^{3/2}\, C_W^3}{K^{1/2}\, (\widetilde{\lambda}^\star_{\min,W})^{3/2}} + \frac{2\sqrt{2}\, q^{3/2}\, C_B^3}{K\, (\widetilde{\lambda}^\star_{\min,B})^{3/2}}\right] \\ &\leq 2\sqrt{2}\, A_{\max}^6 \left[C_W^3 \sqrt{\frac{p^3}{K\, (\widetilde{\lambda}^\star_{\min,W})^3}} + C_B^3 \sqrt{\frac{q^3}{K^2\, (\widetilde{\lambda}^\star_{\min,B})^3}}\right] \\ &\leq 2\sqrt{2}\, A_{\max}^7 \left[C_W^3 \sqrt{\frac{p^3}{N\, (\widetilde{\lambda}^\star_{\min,W})^3}} + C_B^3 \sqrt{\frac{q^3}{N^2\, (\widetilde{\lambda}^\star_{\min,B})^3}}\right].\end{aligned}$$

Thus, there exists a constant $C \coloneqq (2)(58)\sqrt{2}\,\max\{C_W^3, C_B^3\} > 0$, independent of N, p, and q, and a random vector $\boldsymbol{\Delta} \coloneqq I(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}$, such that, for all measurable convex sets $\mathscr{C} \subset \mathbb{R}^{p+q}$, the error of the multivariate normal approximation is bounded above by

$$\begin{aligned}&|\mathbb{P}(I(\boldsymbol{\theta}^\star)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) + \boldsymbol{\Delta} \in \mathscr{C}) - \Phi_d(\mathbf{Z}_d \in \mathscr{C})| \\ \leq\;& 58\, (p+q)^{1/4} \sum_{1 \leq k \leq l \leq K} \mathbb{E}\, \|\boldsymbol{Y}_{k,l}\|_2^3 \\ \leq\;& C\, (p+q)^{1/4}\, A_{\max}^7 \left[\sqrt{\frac{p^3}{(\widetilde{\lambda}^\star_{\min,W})^3\, N}} + \sqrt{\frac{q^3}{(\widetilde{\lambda}^\star_{\min,B})^3\, N^2}}\right].\end{aligned}$$

II. **Demonstrating that $\|I(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}\|_2$ is small with high probability.**

Recall that

$$I(\boldsymbol{\theta}^\star) = \begin{pmatrix} I_W(\boldsymbol{\theta}_W^\star) & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & I_B(\boldsymbol{\theta}_B^\star) \end{pmatrix},$$

where

$$\begin{aligned} I_W(\boldsymbol{\theta}_W^\star) &\coloneqq \sum_{k=1}^K -\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_W}\, \ell_{k,k}(\boldsymbol{\theta}_W^\star, \boldsymbol{X}_{k,k}) \\ I_B(\boldsymbol{\theta}_B^\star) &\coloneqq \sum_{1 \leq k < l \leq K} -\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_B}\, \ell_{k,l}(\boldsymbol{\theta}_B^\star, \boldsymbol{X}_{k,l}), \end{aligned}$$

which implies that

$$I(\boldsymbol{\theta}^\star)^{-1} = \begin{pmatrix} I_W(\boldsymbol{\theta}_W^\star)^{-1} & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & I_B(\boldsymbol{\theta}_B^\star)^{-1} \end{pmatrix}.$$

Using Assumption 2,

$$\lambda_{\min}(I_W(\boldsymbol{\theta}^\star)) = K\, \widetilde{\lambda}^\star_{\min,W} > 0 \tag{34}$$

and

$$\lambda_{\min}(I_B(\boldsymbol{\theta}_B^\star)) \;=\; \binom{K}{2} \widetilde{\lambda}_{\min,B}^\star \;>\; 0. \tag{35}$$

Using (34) and (35), we can bound $\|I(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}\|_2$ by

$$\begin{aligned}
\|I(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}\|_2^2 &= \|I_W(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}_W\|_2^2 + \|I_B(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}_B\|_2^2 \\
&\leq |||I_W(\boldsymbol{\theta}^\star)^{-1/2}|||_2^2\, \|\boldsymbol{R}_W\|_2^2 + |||I_B(\boldsymbol{\theta}^\star)^{-1/2}|||^2\, \|\boldsymbol{R}_B\|_2^2 \\
&\leq \frac{\|\boldsymbol{R}_W\|_2^2}{K\, \widetilde{\lambda}_{\min,W}^\star} + \frac{\|\boldsymbol{R}_B\|_2^2}{\binom{K}{2}\, \widetilde{\lambda}_{\min,B}^\star},
\end{aligned}$$

where $\boldsymbol{R}_W \coloneqq (R_1, \ldots, R_p)$ and $\boldsymbol{R}_B \coloneqq (R_{p+1}, \ldots, R_{p+q})$. As a result,

$$\|I(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}\|_2 \;\leq\; \sqrt{\frac{\|\boldsymbol{R}_W\|_2^2}{K\, \widetilde{\lambda}_{\min,W}^\star} + \frac{\|\boldsymbol{R}_B\|_2^2}{\binom{K}{2}\, \widetilde{\lambda}_{\min,B}^\star}}.$$

Applying Lemma 5.2,

$$\begin{aligned}
\|\boldsymbol{R}_W\|_2^2 &\leq p\, A_{\max}^{12}\, C_W^4\, (C_W + 2)^2\, K^2\, \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1^4 \\
\|\boldsymbol{R}_B\|_2^2 &\leq 4\, q\, A_{\max}^{12}\, C_B^4\, (C_B + 2)^2 \binom{K}{2}^2 \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1^4.
\end{aligned}$$

Using the identity

$$N \;=\; \sum_{k=1}^{K} |\mathcal{A}_k| \;=\; K\, \frac{1}{K} \sum_{k=1}^{K} |\mathcal{A}_k| \;=\; K\, A_{\text{avg}},$$

with the definition $A_{\text{avg}} \coloneqq K^{-1} \sum_{k=1}^{K} |\mathcal{A}_k|$, we have the bound

$$\begin{aligned}
\|I(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}\|_2 &\leq C_3\, A_{\max}^6 \sqrt{\frac{p\, K^2\, \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1^4}{K\, \widetilde{\lambda}_{\min,W}^\star} + \frac{q\, \binom{K}{2}^2\, \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1^4}{\binom{K}{2}\, \widetilde{\lambda}_{\min,B}^\star}} \\
&= C_3\, A_{\max}^6 \sqrt{\frac{p\, K\, \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1^4}{\widetilde{\lambda}_{\min,W}^\star} + \frac{q\, \binom{K}{2}\, \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1^4}{\widetilde{\lambda}_{\min,B}^\star}} \\
&\leq C_3\, A_{\max}^6, \sqrt{\frac{p\, N\, \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1^4}{A_{\text{avg}}\, \widetilde{\lambda}_{\min,W}^\star} + \frac{q\, N^2\, \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1^4}{A_{\text{avg}}^2\, \widetilde{\lambda}_{\min,B}^\star}},
\end{aligned}$$

where $C_3 \coloneqq 2 \max\{C_W^2\, (C_W + 2),\, C_B^2\, (C_B + 2)\} > 0$ is a constant independent of N, p, and q. By Theorem 2.1, there exist constants $C_4 > 0$, $C_5 > 0$, and $N_0 \geq 3$, independent of N, p, and q, such that, for all $N \geq N_0$ and with probability at least $1 - N^{-2}$, the maximum likelihood estimator exists, is unique, and satisfies

$$\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_2 \;\; \leq \;\; C_4 \sqrt{A_{\text{avg}}}\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,W}}}{\widetilde{\lambda}^\epsilon_{\min,W}} \sqrt{\frac{p}{N}}$$

and

$$\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|^2 \;\; \leq \;\; C_5\, A_{\text{avg}}\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,B}}}{\widetilde{\lambda}^\epsilon_{\min,B}} \sqrt{\frac{q}{N^2}}.$$

As a result,

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1 \;\; &\leq \;\; \sqrt{p}\, \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_2 \;\; \leq \;\; C_4 \sqrt{A_{\text{avg}}}\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,W}}}{\widetilde{\lambda}^\epsilon_{\min,W}}\, \frac{p}{\sqrt{N}} \\
\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1 \;\; &\leq \;\; \sqrt{q}\, \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_2 \;\; \leq \;\; C_5\, A_{\text{avg}}\, \frac{\sqrt{\widetilde{\lambda}^\star_{\max,B}}}{\widetilde{\lambda}^\epsilon_{\min,B}}\, \frac{q}{N},
\end{aligned}$$

which leads to the bound

$$\begin{aligned}
\|I(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}\|_2 \;\; &\leq \;\; C_3\, A_{\max}^6 \sqrt{\frac{p\, N\, \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1^4}{A_{\text{avg}}\, \widetilde{\lambda}^\star_{\min,W}} + \frac{q\, N^2\, \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1^4}{A_{\text{avg}}^2\, \widetilde{\lambda}^\star_{\min,B}}} \\
&\leq \;\; C_6\, A_{\max}^6 \sqrt{\frac{A_{\text{avg}}\, (\widetilde{\lambda}^\star_{\max,W})^2\, p^5}{(\widetilde{\lambda}^\epsilon_{\min,W})^4\, \widetilde{\lambda}^\star_{\min,W}\, N} + \frac{A_{\text{avg}}^2\, (\widetilde{\lambda}^\star_{\max,B})^2\, q^5}{(\widetilde{\lambda}^\epsilon_{\min,B})^4\, \widetilde{\lambda}^\star_{\min,B}\, N^2}} \\
&\leq \;\; C_6\, A_{\max}^6 \sqrt{A_{\text{avg}}\, \frac{(\widetilde{\lambda}^\star_{\max,W})^2}{(\widetilde{\lambda}^\epsilon_{\min,W})^5}\, \frac{p^5}{N} + A_{\text{avg}}^2\, \frac{(\widetilde{\lambda}^\star_{\max,B})^2}{(\widetilde{\lambda}^\epsilon_{\min,B})^5}\, \frac{q^5}{N^2}},
\end{aligned}$$

defining $C_6 \coloneqq C_3\, \max\{C_4^4,\, C_5^4\} > 0$. Thus, there exists a constant $C \coloneqq C_6 > 0$, independent of N, p, and q, such that $\boldsymbol{\Delta} \coloneqq I(\boldsymbol{\theta}^\star)^{-1/2}\, \boldsymbol{R}$ satisfies

$$\mathbb{P}\left(\|\boldsymbol{\Delta}\|_2 \leq C\, A_{\max}^6 \sqrt{A_{\text{avg}}\, \frac{(\widetilde{\lambda}^\star_{\max,W})^2}{(\widetilde{\lambda}^\epsilon_{\min,W})^5}\, \frac{p^5}{N} + A_{\text{avg}}^2\, \frac{(\widetilde{\lambda}^\star_{\max,B})^2}{(\widetilde{\lambda}^\epsilon_{\min,B})^5}\, \frac{q^5}{N^2}}\right) \;\; \geq \;\; 1 - \frac{1}{N^2}.$$

Conclusion of proof. We have thus shown—recycling notation of constants—that there exist $C_1 > 0$, $C_2 > 0$, and $N_0 \geq 3$, independent of N, p, and q, and a random vector $\boldsymbol{\Delta} \in \mathbb{R}^{p+q}$ such that, for all integers $N > N_0$ and all measurable convex sets $\mathscr{C} \subset \mathbb{R}^{p+q}$,

$$\begin{aligned}
&|\mathbb{P}(I(\boldsymbol{\theta}^\star)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) + \boldsymbol{\Delta} \in \mathscr{C}) - \Phi_d(\mathbf{Z}_d \in \mathscr{C})| \\
&\leq \;\; C_1\, (p+q)^{1/4}\, A_{\max}^7 \left[\sqrt{\frac{p^3}{(\widetilde{\lambda}^\star_{\min,W})^3\, N}} + \sqrt{\frac{q^3}{(\widetilde{\lambda}^\star_{\min,B})^3\, N^2}}\right],
\end{aligned}$$

where $\boldsymbol{\Delta}$ satisfies

$$\mathbb{P}\left(\|\boldsymbol{\Delta}\|_2 \,\le\, C_2\, A_{\max}^6 \sqrt{A_{\text{avg}}\, \frac{(\widetilde{\lambda}_{\max,W}^{\star})^2}{(\widetilde{\lambda}_{\min,W}^{\epsilon})^5}\, \frac{p^5}{N} + A_{\text{avg}}^2\, \frac{(\widetilde{\lambda}_{\max,B}^{\star})^2}{(\widetilde{\lambda}_{\min,B}^{\epsilon})^5}\, \frac{q^5}{N^2}}\right) \;\ge\; 1 - \frac{1}{N^2}.$$

□

5.1. Auxiliary results for Theorem 2.5

We first recall a theorem due to Raič (2019), restated in Lemma 5.1.

Lemma 5.1 (Theorem 1.1, Raič (2019)). *Consider a sequence of independent random vectors given by $\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots \in \mathbb{R}^p$ with $\mathbb{E}\, \boldsymbol{W}_i = 0$ for all $i \in \{1, 2, \ldots\}$. Define*

$$\boldsymbol{S}_n \;\coloneqq\; \sum_{i=1}^{n} \boldsymbol{W}_i, \qquad n \in \{1, 2, \ldots\},$$

and assume that $\mathbb{V}\, \boldsymbol{S}_n = \boldsymbol{I}_p$. Then, for all measurable convex sets $\mathscr{C} \subset \mathbb{R}^p$,

$$\left|\mathbb{P}(\boldsymbol{S}_n \in \mathscr{C}) - \Phi_p(\boldsymbol{Z}_p \in \mathscr{C})\right| \;\le\; (42\, p^{1/4} + 16) \sum_{i=1}^{n} \mathbb{E}\, \|\boldsymbol{W}_i\|_2^3,$$

where $\boldsymbol{Z}_p$ is a multivariate normal random vector with mean vector $\boldsymbol{0}_p$ and covariance matrix $\boldsymbol{I}_p$ and Φ_p is the corresponding probability distribution.

PROOF OF LEMMA 5.1. The lemma is proved as Theorem 1.1 of Raič (2019). □

5.2. Auxiliary results for Part II in the proof of Theorem 2.5

Lemma 5.2. *Under the assumptions of Theorem 2.5,*

$$\begin{aligned} \|\boldsymbol{R}_W\|_2^2 \;&\le\; p\, A_{\max}^{12}\, C_W^4\, (C_W + 2)^2\, K^2\, \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^{\star}\|_1^4 \\ \|\boldsymbol{R}_B\|_2^2 \;&\le\; 4\, q\, A_{\max}^{12}\, C_B^4\, (C_B + 2)^2 \binom{K}{2}^2 \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^{\star}\|_1^4, \end{aligned}$$

where $\|\boldsymbol{R}_W\|_2^2$ and $\|\boldsymbol{R}_B\|_2^2$ are the normed remainder terms in the proof of Theorem 2.5.

PROOF OF LEMMA 5.2. We bound the remainder terms that arose out of the multivariate Taylor approximation in the proof of Theorem 2.5 using derivatives. Recall that each of the remainder terms R_i ($i \in \{1, \ldots, p+q\}$) in the Lagrange form is given by

$$\begin{aligned} R_i \;&=\; \sum_{j=1}^{p+q} \frac{1}{2} \left[\frac{\partial^2}{\partial\, \theta_j^2} \left[\nabla_{\boldsymbol{\theta}}\, m(\dot{\boldsymbol{\theta}}^{(i)}, \boldsymbol{X})\right]_i\right] (\theta_j - \theta_j^{\star})^2 \\ &\quad + \sum_{1 \le j < r \le p+q} \frac{1}{2} \left[\frac{\partial^2}{\partial\, \theta_j\, \partial\, \theta_r} \left[\nabla_{\boldsymbol{\theta}}\, m(\dot{\boldsymbol{\theta}}^{(i)}, \boldsymbol{X})\right]_i\right] (\theta_j - \theta_j^{\star})\, (\theta_r - \theta_r^{\star}), \end{aligned} \tag{36}$$

where $\dot{\boldsymbol{\theta}}^{(i)} = t_i\,\boldsymbol{\theta} + (1-t_i)\,\boldsymbol{\theta}^\star$ (for some $t_i \in (0,1)$, $i \in \{1,\dots,p+q\}$). If

$$\sup_{\boldsymbol{\theta}\in\mathbb{R}^{p+q}\,:\,\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|_1 \le \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star\|_1} \left|\frac{\partial^2}{\partial\,\theta_j\,\partial\,\theta_r}\,[\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta},\boldsymbol{X})]_i\right| \;\le\; M_i, \qquad 1 \le j \le r \le p,$$

for all $i \in \{1,\dots,p\}$ and

$$\sup_{\boldsymbol{\theta}\in\mathbb{R}^{p+q}\,:\,\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|_1 \le \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star\|_1} \left|\frac{\partial^2}{\partial\,\theta_j\,\partial\,\theta_r}\,[\nabla_{\boldsymbol{\theta}}\, m(\boldsymbol{\theta},\boldsymbol{X})]_i\right| \;\le\; M_i, \qquad 1+p \le j \le r \le p+q,$$

for all $i \in \{1+p,\dots,p+q\}$, then the Lagrange remainders are bounded above by

$$|R_i| \;\le\; \begin{cases} \dfrac{M_i}{2}\,\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1^2 & \text{if } i \in \{1,\dots,p\} \\[2ex] \dfrac{M_i}{2}\,\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1^2 & \text{if } i \in \{p+1,\dots,p+q\} \end{cases}$$

on the set

$$\left\{\boldsymbol{\theta}_W \in \mathbb{R}^p \,:\, \|\boldsymbol{\theta}_W - \boldsymbol{\theta}_W^\star\|_1 \le \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1\right\} \times \left\{\boldsymbol{\theta}_B \in \mathbb{R}^p \,:\, \|\boldsymbol{\theta}_B - \boldsymbol{\theta}_B^\star\|_1 \le \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_B^\star\|_1\right\}.$$

For the rest of the proof, assume that $\boldsymbol{\theta}$ belongs to the above set. By Assumption 2, there exists $C_W > 0$ and $C_B > 0$, independent of N, p, and q, such that

$$\sup_{\boldsymbol{x}_{k,k}\in\mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k})\|_\infty \;\le\; C_W \binom{|\mathcal{A}_k|}{2}, \qquad 1 \le k \le K,$$

and

$$\sup_{\boldsymbol{x}_{k,l}\in\mathbb{X}_{k,l}} \|s_{k,l}(\boldsymbol{x}_{k,l})\|_\infty \;\le\; C_B\,|\mathcal{A}_k|\,|\mathcal{A}_l|, \qquad 1 \le k < l \le K,$$

Lemmas 5.3 and 5.4 establish that

$$\left|\frac{\partial^2}{\partial\,\theta_j\,\partial\,\theta_h}\,[\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{\theta},\boldsymbol{X})]_i\right| \;\le\; \begin{cases} A_{\max}^6\,C_W^3\,(C_W+2)\,K, & (i,j,h) \in \{1,\dots,p\}^3 \\ 2\,A_{\max}^6\,C_B^2\,(C_B+2)\,\binom{K}{2}, & (i,j,h) \in \{p+1,\dots,p+q\}^3 \\ 0 & \text{otherwise} \end{cases}.$$

As a result, when $m(\boldsymbol{\theta},\boldsymbol{X}) = \ell(\boldsymbol{\theta},\boldsymbol{X})$ in the proof of Theorem 2.5,

$$|R_i| \;\le\; \begin{cases} A_{\max}^6\,C_W^3\,(C_W+2)\,K\,\|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1^2, & 1 \le i \le p \\[2ex] 2\,A_{\max}^6\,C_B^2\,(C_B+2)\,\binom{K}{2}\,\|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1^2, & p+1 \le i \le p+q \end{cases},$$

which implies the bounds

$$\begin{aligned} \|\boldsymbol{R}_W\|_2^2 &\le \sum_{i=1}^{p} R_i^2 &\le\; p\, A_{\max}^{12}\, C_W^4\, (C_W+2)^2\, K^2\, \|\widehat{\boldsymbol{\theta}}_W - \boldsymbol{\theta}_W^\star\|_1^4 \\ \|\boldsymbol{R}_B\|_2^2 &\le \sum_{i=p+1}^{p+q} R_i^2 &\le\; 4\, q\, A_{\max}^{12}\, C_B^4\, (C_B+2)^2 \binom{K}{2}^2 \|\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_B^\star\|_1^4. \end{aligned}$$

□

Lemma 5.3. *Consider an exponential-family local dependence random graph model which satisfies Assumption 1. Then, for all $(i, j, h) \in \{1, \ldots, p\}^3$,*

$$\left|\frac{\partial^2\, [\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{x})]_i}{\partial \theta_h\, \partial \theta_j}\right| \;\le\; A_{\max}^6\, C_W^2\, (C_W + 2).$$

Proof of Lemma 5.3. By Lemma 6.1, the second derivatives of the log-likelihood taken with respect to the natural parameters are equal to the variances and covariances of the sufficient statistics of the exponential family, implying, for all $(i, j) \in \{1, \ldots, p\}^2$, that

$$\frac{\partial^2\, \ell(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \theta_j\, \partial \theta_i} \;=\; \frac{\partial\, [\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{x})]_i}{\partial \theta_j} \;=\; \mathbb{C}_{\boldsymbol{\theta}}\left(\sum_{k=1}^{K} s_{k,k,i}(\boldsymbol{X}_{k,k}),\; \sum_{k=1}^{K} s_{k,k,j}(\boldsymbol{X}_{k,k})\right),$$

where $\mathbb{C}_{\boldsymbol{\theta}}$ denotes the covariance operator corresponding to the probability distribution $\mathbb{P}_{\boldsymbol{\theta}}$. By the independence of the block-based subgraphs $\boldsymbol{X}_{k,k}$ $(1 \le k \le K)$,

$$\mathbb{C}_{\boldsymbol{\theta}}\left(\sum_{k=1}^{K} s_{k,k,i}(\boldsymbol{X}_{k,k}),\; \sum_{k=1}^{K} s_{k,k,j}(\boldsymbol{X}_{k,k})\right) \;=\; \sum_{k=1}^{K} \mathbb{C}_{\boldsymbol{\theta}}(s_{k,k,i}(\boldsymbol{X}_{k,k}),\; s_{k,k,j}(\boldsymbol{X}_{k,k})).$$

Taking $h \in \{1, \ldots, p\}$ and using the triangle inequality, we obtain the bound

$$\left|\frac{\partial^2\, [\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{x})]_i}{\partial \theta_h\, \partial \theta_j}\right| \;\le\; \sum_{k=1}^{K} \left|\frac{\partial}{\partial\, \theta_h}\, \mathbb{C}_{\boldsymbol{\theta}}(s_{k,k,i}(\boldsymbol{X}_{k,k}),\; s_{k,k,j}(\boldsymbol{X}_{k,k}))\right|. \tag{37}$$

To proceed from here, we apply Lemma 6.3. To do so, we verify the assumptions of Lemma 6.3. By Assumption 1, there exists $C_W > 0$ such that

$$\sup_{\boldsymbol{x}_{k,k} \in \mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k})\|_\infty \;\le\; C_W \binom{|\mathcal{A}_k|}{2} \;\le\; \frac{C_W}{2}\, A_{\max}^2, \qquad k \in \{1, \ldots, K\},$$

which implies, for all $k \in \{1, \ldots, K\}$, that

$$\sup_{\boldsymbol{x}_{k,k} \in \mathbb{X}_{k,k}} \|s_{k,k}(\boldsymbol{x}_{k,k}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_{k,k}(\boldsymbol{X}_{k,k})\|_\infty \;\le\; 2\, C_W \binom{|\mathcal{A}_k|}{2} \;\le\; C_W\, A_{\max}^2.$$

Taking $U_1 = (C_W\,/\,2)\,A_{\max}^2 > 0$ and $U_2 = C_W\,A_{\max}^2 > 0$ and applying Lemma 6.3,

$$\begin{aligned}\left|\frac{\partial}{\partial\,\theta_h}\,\mathbb{C}_{\boldsymbol{\theta}}(s_{k,k,i}(\boldsymbol{X}_{k,k}),\, s_{k,k,j}(\boldsymbol{X}_{k,k}))\right| \;&\leq\; \left(C_W\,A_{\max}^2\right)\left(C_W\,A_{\max}^2\right)\left(C_W\,A_{\max}^2+2\right)\\ &\leq\; A_{\max}^6\,C_W^2\,(C_W+2).\end{aligned}$$

Hence, for all $\{i, j, h\} \subseteq \{1, \ldots, p\}$,

$$\left|\frac{\partial^2\,[\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{\theta},\boldsymbol{x})]_i}{\partial\theta_h\,\partial\theta_j}\right| \;\leq\; \sum_{k=1}^{K} A_{\max}^6\,C_W^2\,(C_W+2) \;\leq\; A_{\max}^6\,C_W^2\,(C_W+2)\,K.$$

□

Lemma 5.4. *Consider an exponential-family local dependence random graph model which satisfies Assumption 1. Then, for all* $(i, j, h) \in \{p+1, \ldots, p+q\}^3$,

$$\left|\frac{\partial^2\,[\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{\theta};\boldsymbol{x})]_i}{\partial\theta_h\,\partial\theta_j}\right| \;\leq\; 2\,A_{\max}^6\,C_B^2\,(C_B+2)\binom{K}{2}.$$

PROOF OF LEMMA 5.4. Lemma 5.4 is proved similarly to Lemma 5.3, with the notable exception that the sum in (37) is over the index set $1 \leq k < l \leq K$, for the between-block subgraphs. As a result, the factor of K in the bound in Lemma 5.3 is replaced with $\binom{K}{2}$. The bound $|\mathcal{A}_k|\,|\mathcal{A}_l| \leq A_{\max}^2$ is used in place of $\binom{|\mathcal{A}_k|}{2} \leq A_{\max}^2$, resulting in an extra factor of 2. The rest of the proof can be repeated unchanged, with the appropriate adjustments to indexing (e.g., using $C_B > 0$ in place of $C_W > 0$). □

Lemma 5.5. *Let* $a_1, a_2, b_1, b_2 \in \mathbb{R}$. *Then*

$$|a_1\,b_1 - a_2\,b_2| \;\leq\; |a_1|\,|b_1 - b_2| + |b_2|\,|a_1 - a_2|.$$

PROOF OF LEMMA 5.5. Write

$$\begin{aligned}|a_1\,b_1 - a_2\,b_2| \;&=\; |a_1\,b_1 - (a_2 - a_1 + a_1)\,b_2|\\ &=\; |a_1\,b_1 - b_2\,(a_2 - a_1) - a_1\,b_2|\\ &=\; |a_1\,(b_1 - b_2) - b_2\,(a_2 - a_1)|\\ &\leq\; |a_1|\,|b_1 - b_2| + |b_2|\,|a_1 - a_2|.\end{aligned}$$

□

6. Auxiliary results for exponential families

Lemma 6.1. *Consider a random vector $\boldsymbol{Y}$ with finite support $\mathbb{Y}$ (i.e., $|\mathbb{Y}| < \infty$) and assume that the probability mass function $f_{\boldsymbol{\theta}} : \mathbb{Y} \mapsto (0,1)$ belongs to an m-dimensional exponential family, i.e.,*

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) \;=\; h(\boldsymbol{y}) \exp\left(\langle \boldsymbol{\theta},\, s(\boldsymbol{y})\rangle - \psi(\boldsymbol{\theta})\right), \qquad \boldsymbol{y} \in \mathbb{Y},\ \boldsymbol{\theta} \in \mathbb{R}^m.$$

Then

$$\begin{aligned} \nabla_{\boldsymbol{\theta}}\, \psi(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{Y}) \\ \nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{y}) &= s(\boldsymbol{y}) - \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{Y}) \\ \nabla^2_{\boldsymbol{\theta}}\, \psi(\boldsymbol{\theta}) &= -\nabla^2_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{y}) \;=\; \mathbb{V}_{\boldsymbol{\theta}}\, s(\boldsymbol{Y}). \end{aligned}$$

PROOF OF LEMMA 6.1. All results follow from Propositions 3.8 and 3.10 of Sundberg (2019). □

Lemma 6.2. *Consider a random vector $\boldsymbol{Y}$ with finite support $\mathbb{Y}$ (i.e., $|\mathbb{Y}| < \infty$) and assume that the probability mass function $f_{\boldsymbol{\theta}} : \mathbb{Y} \mapsto (0,1)$ belongs to an m-dimensional exponential family, i.e.,*

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) \;=\; h(\boldsymbol{y}) \exp\left(\langle \boldsymbol{\theta},\, s(\boldsymbol{y})\rangle - \psi(\boldsymbol{\theta})\right), \qquad \boldsymbol{y} \in \mathbb{Y},\ \boldsymbol{\theta} \in \mathbb{R}^m.$$

Then

$$\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{Y}) - \mathbb{E}\, \nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{Y}) \;=\; s(\boldsymbol{Y}) - \mathbb{E}\, s(\boldsymbol{Y}), \qquad \boldsymbol{\theta} \in \mathbb{R}^m,$$

and

$$\sup_{\boldsymbol{\theta} \in \mathbb{R}^m} \|\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{Y}) - \mathbb{E}\, \nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{Y})\|_\infty \;=\; \|s(\boldsymbol{Y}) - \mathbb{E}\, s(\boldsymbol{Y})\|_\infty \;=\; \|\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}^\star, \boldsymbol{Y})\|_\infty.$$

PROOF OF LEMMA 6.2. Applying Lemma 6.1, $\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{y}) = s(\boldsymbol{y}) - \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{Y})$. Hence, for all $\boldsymbol{\theta} \in \mathbb{R}^m$,

$$\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{Y}) - \mathbb{E}\, \nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{Y}) \;=\; s(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{Y}) - \mathbb{E}\, s(\boldsymbol{Y}) + \mathbb{E}\, \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{Y}) \;=\; s(\boldsymbol{Y}) - \mathbb{E}\, s(\boldsymbol{Y}),$$

which implies the final result

$$\sup_{\boldsymbol{\theta} \in \mathbb{R}^m} \|\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{Y}) - \mathbb{E}\, \nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}, \boldsymbol{Y})\|_\infty \;=\; \sup_{\boldsymbol{\theta} \in \mathbb{R}^m} \|s(\boldsymbol{Y}) - \mathbb{E}\, s(\boldsymbol{Y})\|_\infty \;=\; \|s(\boldsymbol{Y}) - \mathbb{E}\, s(\boldsymbol{Y})\|_\infty.$$

□

Lemma 6.3. *Let $\boldsymbol{Y}$ be a random vector with finite support $\mathbb{Y}$ (i.e., $|\mathbb{Y}| < \infty$) and assume that the distribution of $\boldsymbol{Y}$ belongs to an exponential family with with probability mass functions of the form*

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) \;=\; h(\boldsymbol{y}) \exp(\langle \boldsymbol{\theta},\, \boldsymbol{y}\rangle - \psi(\boldsymbol{\theta})), \qquad \boldsymbol{y} \in \mathbb{Y},\ \boldsymbol{\theta} \in \mathbb{R}^m.$$

Assume that there exist constants $U_1 > 0$ and $U_2 > 0$ such that, for all $t \in \{1, \ldots, m\}$,

$$|Y_t| \;\leq\; U_1 \qquad \text{and} \qquad |Y_t - \mathbb{E}_{\boldsymbol{\theta}}\, Y_t| \;\leq\; U_2, \qquad \mathbb{P}\text{-almost surely}.$$

Then, for all $(i, j, h) \in \{1, \ldots, m\}^3$,

$$\left|\frac{\partial}{\partial\, \theta_h}\, \mathbb{C}_{\boldsymbol{\theta}}(Y_i,\, Y_j)\right| \;\leq\; 2\, U_1\, U_2\, (U_2 + 2).$$

PROOF OF LEMMA 6.3. Let $(i, j, h) \in \{1, \dots, m\}^3$ and define $\mu_t(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\boldsymbol{\theta}} Y_t$ ($t \in \{1, \dots, m\}$. Then

$$\begin{aligned}
\frac{\partial}{\partial \theta_h} \mathbb{C}_{\boldsymbol{\theta}}(Y_i, Y_j) &= \frac{\partial}{\partial \theta_h} \mathbb{E}_{\boldsymbol{\theta}}[Y_i Y_j - \mu_i(\boldsymbol{\theta}) \mu_j(\boldsymbol{\theta})] \\
&= \frac{\partial}{\partial \theta_h} \sum_{\boldsymbol{y} \in \mathbb{Y}} \left[y_i y_j - \mu_i(\boldsymbol{\theta}) \mu_j(\boldsymbol{\theta})\right] f_{\boldsymbol{\theta}}(\boldsymbol{y}) \\
&= \sum_{\boldsymbol{y} \in \mathbb{Y}} \left[f_{\boldsymbol{\theta}}(\boldsymbol{y}) \frac{\partial}{\partial \theta_h} \left[y_i y_j - \mu_i(\boldsymbol{\theta}) \mu_j(\boldsymbol{\theta})\right] + \left[y_i y_j - \mu_i(\boldsymbol{\theta}) \mu_j(\boldsymbol{\theta})\right] \frac{\partial}{\partial \theta_h} f_{\boldsymbol{\theta}}(\boldsymbol{y})\right].
\end{aligned}$$

Using Lemma 6.1 and applying the chain rule,

$$\frac{\partial}{\partial \theta_h} f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{\partial}{\partial \theta_h} h(\boldsymbol{y}) \exp(\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle - \psi(\boldsymbol{\theta})) = [y_h - \mu_h(\boldsymbol{\theta})] f_{\boldsymbol{\theta}}(\boldsymbol{y}).$$

Hence,

$$\frac{\partial}{\partial \theta_h} \mathbb{C}_{\boldsymbol{\theta}}(Y_i, Y_j) = \mathbb{E}_{\boldsymbol{\theta}}[(Y_i Y_j - \mu_i(\boldsymbol{\theta}) \mu_j(\boldsymbol{\theta}))(Y_h - \mu_h(\boldsymbol{\theta}))] - \frac{\partial}{\partial \theta_h} \left[\mu_i(\boldsymbol{\theta}) \mu_j(\boldsymbol{\theta})\right].$$

We next compute, using Lemma 6.1,

$$\begin{aligned}
\frac{\partial}{\partial \theta_h} \left[\mu_i(\boldsymbol{\theta}) \mu_j(\boldsymbol{\theta})\right] &= \mu_i(\boldsymbol{\theta}) \mathbb{C}_{\boldsymbol{\theta}}(Y_j, Y_h) + \mu_j(\boldsymbol{\theta}) \mathbb{C}_{\boldsymbol{\theta}}(Y_i, Y_h) \\
&= \mu_i(\boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\theta}}[Y_j Y_h - \mu_j(\boldsymbol{\theta}) \mu_h(\boldsymbol{\theta})] + \mu_j(\boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\theta}}[Y_i Y_h - \mu_i(\boldsymbol{\theta}) \mu_h(\boldsymbol{\theta})].
\end{aligned}$$

By Jensen's inequality and the triangle inequality

$$\begin{aligned}
\left|\frac{\partial}{\partial \theta_h} \mathbb{C}_{\boldsymbol{\theta}}(Y_i, Y_j)\right| \leq\; & \mathbb{E}_{\boldsymbol{\theta}} \left[\left|(Y_i Y_j - \mu_i(\boldsymbol{\theta}) \mu_j(\boldsymbol{\theta}))\right| |(Y_h - \mu_h(\boldsymbol{\theta}))|\right] \\
&+ |\mu_i(\boldsymbol{\theta})| \mathbb{E}_{\boldsymbol{\theta}} \left|Y_j Y_h - \mu_j(\boldsymbol{\theta}) \mu_h(\boldsymbol{\theta})\right| \\
&+ |\mu_j(\boldsymbol{\theta})| \mathbb{E}_{\boldsymbol{\theta}} |Y_i Y_h - \mu_i(\boldsymbol{\theta}) \mu_h(\boldsymbol{\theta})|.
\end{aligned}$$

The assumption there exist constants $U_1 > 0$ and $U_2 > 0$ such that $|Y_t| \leq U_1$ for all $t \in \{1, \dots, m\}$ and $|Y_t - \mu_t(\boldsymbol{\theta})| \leq U_2$ ($t \in \{1, \dots, m\}$) hold $\mathbb{P}$-almost surely implies that $|\mu_t(\boldsymbol{\theta})| \leq U_1$ for all $t \in \{1, \dots, m\}$, and, through an application of Lemma 5.5, that

$$\left|Y_j Y_h - \mu_j(\boldsymbol{\theta}) \mu_h(\boldsymbol{\theta})\right| \leq |Y_j| |Y_h - \mu_h(\boldsymbol{\theta})| + |\mu_h(\boldsymbol{\theta})| |Y_j - \mu_j(\boldsymbol{\theta})| \leq 2 U_1 U_2.$$

Hence,

$$\left|\frac{\partial}{\partial \theta_h} \mathbb{C}_{\boldsymbol{\theta}}(Y_i, Y_j)\right| \leq 2 U_1 U_2^2 + 4 U_1 U_2 = 2 U_1 U_2 (U_2 + 2).$$

□

Lemma 6.4. *Consider a random vector $\boldsymbol{Y}$ with finite support $\mathbb{Y}$ (i.e., $|\mathbb{Y}| < \infty$) and assume that the distribution of $\boldsymbol{Y}$ belongs to an exponential family with probability mass functions of the form*

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}) \;=\; h(\boldsymbol{y}) \exp\left(\langle \boldsymbol{\theta}, s(\boldsymbol{y})\rangle - \psi(\boldsymbol{\theta})\right), \qquad \boldsymbol{y} \in \mathbb{Y},\ \boldsymbol{\theta} \in \mathbb{R}^m.$$

Then for all functions $f : \mathbb{Y} \mapsto \mathbb{R}$,

$$\frac{\partial}{\partial \theta_i} \mathbb{E}_{\boldsymbol{\theta}} f(\boldsymbol{Y}) \;=\; \mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{Y})\,(s_i(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_i(\boldsymbol{Y}))],$$

for all $i \in \{1, \ldots, m\}$.

PROOF OF LEMMA 6.4. Write

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \mathbb{E}_{\boldsymbol{\theta}} f(\boldsymbol{Y}) &= \frac{\partial}{\partial \theta_i} \sum_{\boldsymbol{y} \in \mathbb{Y}} f(\boldsymbol{y})\, h(\boldsymbol{y}) \exp\left(\langle \boldsymbol{\theta}, s(\boldsymbol{y})\rangle - \psi(\boldsymbol{\theta})\right) \\
&= \sum_{\boldsymbol{y} \in \mathbb{Y}} f(\boldsymbol{y})\, h(\boldsymbol{y}) \left[\frac{\partial}{\partial \theta_i} \exp\left(\langle \boldsymbol{\theta}, s(\boldsymbol{y})\rangle - \psi(\boldsymbol{\theta})\right)\right] \\
&= \sum_{\boldsymbol{y} \in \mathbb{Y}} f(\boldsymbol{y})\, h(\boldsymbol{y}) \exp\left(\langle \boldsymbol{\theta}, s(\boldsymbol{y})\rangle - \psi(\boldsymbol{\theta})\right)\, (s_i(\boldsymbol{y}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_i(\boldsymbol{Y})) \\
&= \mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{Y})\,(s_i(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\theta}}\, s_i(\boldsymbol{Y}))],
\end{aligned}$$

as applying Lemma 6.1 shows that

$$\frac{\partial}{\partial \theta_i} \psi(\boldsymbol{\theta}) \;=\; \mathbb{E}_{\boldsymbol{\theta}}\, s_i(\boldsymbol{X}), \qquad i = 1, \ldots, m.$$

□

References

BROWN, L. D. (1986). *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics.

CHEN, P., GAO, C. and ZHANG, A. Y. (2022). Partial recovery for top-k ranking: optimality of MLE and sub-optimality of spectral method. *Ann. Statist.* **50** 1618-1652.

ORTEGA, J. M. and RHEINBOLDT, W. C. (2000). *Iterative solution of nonlinear equations in several variables*. Society for Industrial and Applied Mathematics.

RAIČ, M. (2019). A multivariate Berry-Esseen theorem with explicit constants. *Bernoulli* **25** 2824-2853.

SUNDBERG, R. (2019). *Statistical Modelling by Exponential Families*. Cambridge University Press.

VERSHYNIN, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge university press.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge university press.