Statistical Inference of Random Graphs With a Surrogate Likelihood Function

Dingbo Wu   DW16@IU.EDU
Department of Statistics
Indiana University
Bloomington, IN 47405, USA

Fangzheng Xie   FXIE@IU.EDU
Department of Statistics
Indiana University
Bloomington, IN 47405, USA

Editor: Rajarshi Guhaniyogi

Abstract

Spectral estimators have been broadly applied to statistical network analysis, but they do not incorporate the likelihood information of the network sampling model. This paper proposes a novel surrogate likelihood function for statistical inference of a class of popular network models referred to as random dot product graphs. In contrast to the structurally complicated exact likelihood function, the surrogate likelihood function has a separable structure and is log-concave yet approximates the exact likelihood function well. From the frequentist perspective, we study the maximum surrogate likelihood estimator and establish the accompanying theory. We show its existence, uniqueness, large sample properties, and that it improves upon the baseline spectral estimator with a smaller sum of squared errors. Furthermore, we derive the second-order bias of the proposed estimator and gain insight into why it outperforms some of the existing estimators. A computationally convenient stochastic gradient descent algorithm is designed to find the maximum surrogate likelihood estimator in practice. From the Bayesian perspective, we establish the Bernstein–von Mises theorem of the posterior distribution with the surrogate likelihood function and show that the resulting credible sets have the correct frequentist coverage. The empirical performance of the proposed surrogate-likelihood-based methods is validated through the analyses of simulation examples and two real-world data sets.

Keywords: Bernstein–von Mises theorem, maximum surrogate likelihood estimation, random dot product graph, stochastic gradient descent

1. Introduction

In the contemporary world of data science, network data are pervasive in a broad range of applications such as sociology (Lacetera et al., 2016; Young and Scheinerman, 2007), econometrics (Mele, 2017; Mele et al., 2022), and neuroscience (Tang et al., 2019). Statistical network analysis is also an interdisciplinary area of research connected with many other fields, including computer science, machine learning, combinatorics, applied mathematics, and physics. To model and analyze network data, various random graph models have

been proposed in the literature, including the Erdös-Rényi graph (Erdős et al., 1960), the stochastic block model (Holland et al., 1983), and the latent space model (Hoff et al., 2002).

In this paper, we focus on random dot product graphs (Young and Scheinerman, 2007), a class of random graph models that are popular due to its simple architecture and flexibility. On one hand, the edge probability matrix of a random dot product graph has a low-rank structure, which motivates, among others, the use of spectral methods in statistical network analysis. On the other hand, the random dot product graph model is quite flexible because it not only encompasses the popularly used stochastic block models (Holland et al., 1983; Abbe et al., 2016) and their offspring (Airoldi et al., 2008; Binkiewicz et al., 2017; Lyzinski et al., 2017; Sengupta and Chen, 2018), but can also approximate general latent position graphs when the rank of the edge probability matrix grows with the number of vertices at a certain rate (Gao et al., 2015; Tang et al., 2013).

Because the adjacency matrix has a low expected rank, spectral quantities such as the leading eigenvectors of the adjacency matrix and those of the normalized Laplacian matrix, have been extensively used for low-rank random graph inference. In particular, it is well known that the rows of these eigenvectors encode the cluster membership information when the underlying graph is generated from a stochastic block model (Abbe et al., 2020; Lyzinski et al., 2014; Lei and Rinaldo, 2015; Rohe et al., 2011; Sussman et al., 2012). There has been substantial recent development on the theory for spectral methods and the corresponding subsequent inference tasks in random dot product graphs. For an incomplete list of reference, see Athreya et al. (2016); Sussman et al. (2014); Sarkar and Bickel (2015); Tang and Priebe (2018); Tang et al. (2013, 2017a,b). The readers are also referred to the survey paper Athreya et al. (2017) for a review of the recent advances in this topic.

It has been pointed out (Xie and Xu, 2020, 2023; Xie, 2024) that, although the spectral methods for random dot product graphs have gained marvelous success and been broadly applied, the Bernoulli likelihood information contained in the graph distribution has been neglected. This observation has motivated the development of likelihood-based inference for random dot product graphs. Xie and Xu (2020) proposed a fully Bayesian approach for estimating the latent positions in random dot product graphs, referred to as posterior spectral embedding, and established its global minimax optimality. Li et al. (2023) studied the maximum likelihood estimation for a general class of latent space networks and established the asymptotic normality of the resulting estimator under a slightly different setup. Xie and Xu (2023) proposed a novel one-step procedure, which lead to a one-step estimator that took advantage of the Bernoulli likelihood information of the sampling model through the score function and the Fisher information matrix, to estimate random dot product graphs from the frequentist perspective. There, the authors further established the asymptotic efficiency of the one-step estimator and its smaller asymptotic sum of squared errors compared to that of the spectral estimators. The sparsity condition imposed in Xie and Xu (2023) was significantly weakened by Xie (2024) through a cleverly-designed leave-one-out analysis and delicate concentration analyses. Later, Tang et al. (2022) applied the idea of the one-step refinement of spectral methods to stochastic block models when the block probability matrix is rank deficient.

Despite the success of the one-step estimator, a central question regarding likelihood-based inference for random dot product graphs remains open: What is the behavior of the frequentist maximum likelihood estimator? Also, a related question is: What is the behav-

ior of the Bayes estimator? Efforts attempting to address these two questions aim to gain deeper insight into the likelihood-based inference for random dot product graphs from the frequentist and the Bayesian perspective, respectively. These two questions are also closely related through the Bernstein–von Mises phenomenon (see, for example, Section 10.2 in Van der Vaart, 2000). Here, the major technical barrier is the complicated structure of the parameter space for the latent positions. In this paper, we partially answer the two aforementioned questions by resorting to a cleverly-designed surrogate likelihood function that simplifies the parameter space enormously. Our work features the following novel contributions: Firstly, the surrogate likelihood function has a separable structure, is log-concave, and the associated parameter space for the latent positions is a convex relaxation of the original latent space. These features greatly facilitate both the theoretical analysis and the related practical computation. Secondly, we establish the existence, uniqueness, and asymptotic efficiency of the frequentist maximum surrogate likelihood estimator under the minimal sparsity condition. In particular, similar to the one-step estimator, the maximum surrogate likelihood estimator improves upon the baseline spectral estimators with a smaller sum of squared errors. Furthermore, we have rigorously derived the second-order bias formulae of the maximum surrogate likelihood estimator and the one-step estimator, thereby providing insight into why the former typically outperforms the latter in certain finite sample problems. Thirdly, we design a computationally efficient stochastic gradient descent algorithm for the maximum surrogate likelihood estimator with adaptive step sizes. Fourthly, regarding the Bayes procedure, we establish the Bernstein–von Mises theorem for the posterior distribution with the surrogate likelihood function and show that the resulting credible sets have the correct frequentist coverage probabilities.

The remaining part of the article is structured as follows. In Section 2, we review the background of random dot product graphs and introduce the surrogate likelihood function. Section 3 elaborates on the theoretical properties and the computational algorithm of the maximum surrogate likelihood estimation. In Section 4, we establish the large sample properties of the Bayes procedure with the surrogate likelihood function. Section 5 demonstrates the empirical performance of the proposed methods through simulation examples and the analysis of a real-world Wikipedia network data set. We conclude the paper with a discussion in Section 6.

Notations: Let $[n]$ denote the set of consecutive integers from 1 to n: $[n] = \{1, \ldots, n\}$. The symbol $\lesssim_\delta$ means an inequality up to a constant depending on δ, that is, $a \lesssim_\delta b$ if $a \le C_\delta b$ for some constant $C_\delta > 0$ depending on δ; a similar definition also applies to the symbol $\gtrsim_\delta$. The notation $\|\mathbf{x}\|$ denotes the Euclidean norm of a vector $\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{T}} \in \mathbb{R}^d$, that is, $\|\mathbf{x}\| = (\sum_{k=1}^d x_k^2)^{1/2}$. The $d \times d$ identity matrix is denoted by $\mathbf{I}_d$. The notation $\mathbb{O}(n, d) = \{\mathbf{U} \in \mathbb{R}^{n \times d} : \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_d\}$ denotes the set of all orthonormal d-frames in $\mathbb{R}^n$, where $d \le n$, and we write $\mathbb{O}(d) = \mathbb{O}(d, d)$. For a matrix $\mathbf{X} = [x_{ik}]_{n \times d}$, $\sigma_k(\mathbf{X})$ denotes its kth largest singular value, and when $\mathbf{X}$ is square and symmetric, $\lambda_k(\mathbf{X})$ denotes its kth largest eigenvalue in magnitude. Matrix norms with following definitions are used: the spectral norm $\|\mathbf{X}\|_2 = \sigma_1(\mathbf{X})$, the Frobenius norm $\|\mathbf{X}\|_{\mathrm{F}} = (\sum_{i=1}^n \sum_{k=1}^d x_{ik}^2)^{1/2}$, the matrix infinity norm $\|\mathbf{X}\|_\infty = \max_{i \in [n]} \sum_{k=1}^d |x_{ik}|$, and the two-to-infinity norm $\|\mathbf{X}\|_{2 \to \infty} = \max_{i \in [n]} (\sum_{k=1}^d x_{ik}^2)^{1/2}$. In particular, these norm notations apply to any Euclidean vector $\mathbf{x} \in \mathbb{R}^d$ viewed as a $d \times 1$ matrix. Given two symmetric positive semidefinite matrices $\mathbf{A}, \mathbf{B}$

3

of the same dimension, we write $\mathbf{A} \preceq \mathbf{B}$ ($\mathbf{A} \succeq \mathbf{B}$, respectively) if $\mathbf{B} - \mathbf{A}$ ($\mathbf{A} - \mathbf{B}$, respectively) is positive semidefinite.

2. Background and the Surrogate Likelihood

In this section, we review the background of random dot product graphs and introduce the surrogate likelihood function.

2.1 Background on Random Dot Product Graphs

We begin by briefly reviewing the background on random dot product graphs and adjacency spectral embedding. Consider a graph with n vertices labeled as $[n] = \{1, \ldots, n\}$. Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ such that $\mathbf{x}_1^{\mathrm{T}} \mathbf{x}_2 \in (0, 1)$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, where d is fixed and $d \leq n$, and let $\rho_n \in (0, 1]$ be a sparsity factor. Each vertex $i \in [n]$ is associated with a vector $\mathbf{x}_i \in \mathcal{X}$, referred to as the latent position for vertex i. We say that a symmetric random matrix $\mathbf{A} = [A_{ij}]_{n \times n} \in \{0, 1\}^{n \times n}$ is an adjacency matrix generated by a random dot product graph with latent position matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}}$ and sparsity factor ρ_n, denoted by $\mathbf{A} \sim \mathrm{RDPG}(\rho_n^{1/2} \mathbf{X})$, if the random variables $A_{ij} \sim \mathrm{Bernoulli}(\rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j)$ independently for all $i, j \in [n]$, $i \leq j$, and $A_{ij} = A_{ji}$ for all $i > j$. The distribution of $\mathbf{A}$ can thus be written as $p_{\mathbf{X}}(\mathbf{A}) = \prod_{i \leq j} (\rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j)^{A_{ij}} (1 - \rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j)^{1 - A_{ij}}$. The sparsity factor ρ_n fundamentally controls the overall average graph expected degree through $n \rho_n$ when the entries of $\mathbf{X} \mathbf{X}^{\mathrm{T}}$ are bounded away from 0 and ∞.

Remark 1 (Deterministic versus stochastic latent positions) *In this work, we consider the setup where the latent positions $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are deterministic parameters to be estimated. Another slightly different modeling approach is to consider $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as independent and identically distributed latent random variables (see, for example, Athreya et al., 2016; Tang et al., 2017b; Tang and Priebe, 2018). This random formulation of the latent positions introduces implicit homogeneity and is connected to the infinite exchangeable random graphs (Janson and Diaconis, 2008). The same homogeneity condition was retained in Xie and Xu (2023) using a Glivenko–Cantelli type condition when $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are deterministic. The latter Glivenko–Cantelli type condition is also relaxed in the current work as we only require that $\sigma_d(\mathbf{X}) > 0$ (see Remark 2 below).*

Remark 2 (Nonidentifiability) *The latent position matrix $\mathbf{X}$ is not uniquely identified in the following two senses. Firstly, any low-rank positive semidefinite connection probability matrix $\mathbf{P} = \mathbf{X} \mathbf{X}^{\mathrm{T}}$ can have different factorizations because for any orthogonal matrix $\mathbf{W} \in \mathbb{O}(d)$, $\mathbf{X} \mathbf{X}^{\mathrm{T}} = (\mathbf{X} \mathbf{W})(\mathbf{X} \mathbf{W})^{\mathrm{T}}$. Secondly, for any $d' > d$ and any latent position matrix $\mathbf{X} \in \mathbb{R}^d$, there exists another matrix $\mathbf{X}' \in \mathbb{R}^{d'}$ such that $\mathbf{X} \mathbf{X}^{\mathrm{T}} = \mathbf{X}'(\mathbf{X}')^{\mathrm{T}}$. The latter source of non-identifiability can be removed by requiring that $\sigma_d(\mathbf{X}) > 0$, while the former source is inevitable without further constraints. Thus, any estimator of the latent position matrix $\mathbf{X}$ can only recover it up to an orthogonal transformation.*

Example 1 (Stochastic block model) *Random dot product graphs have connections with the popular stochastic block model (Holland et al., 1983). Consider a graph with n vertices that are partitioned into K communities, where K is assumed to be much smaller than n.*

Let $\tau : [n] \to [K]$ be a cluster assignment function that assigns each vertex to a unique community. Let $\mathbf{B} = [B_{kl}]_{K \times K} \in (0,1)^{K \times K}$ be a symmetric probability matrix and A_{ij} be the binary indicator of the existence of an edge between vertices i and j. Then the stochastic block model specifies that $A_{ij} \sim \text{Bernoulli}(B_{\tau(i)\tau(j)})$ independently for all $i, j \in [n]$, $i \leq j$, and $A_{ij} = A_{ji}$ for all $i > j$. By converting the community assignment to a matrix $\mathbf{Z} = [1\{\tau(i) = k\}]_{n \times K}$, we see that the expected adjacency matrix $\mathbf{Z}\mathbf{B}\mathbf{Z}^{\mathrm{T}}$ is symmetric and of low rank. Furthermore, if $\mathbf{B}$ is positive semidefinite with rank $d \leq K$ and can be factorized as $\mathbf{B} = \mathbf{V}\mathbf{V}^{\mathrm{T}}$ for a $K \times d$ matrix $\mathbf{V}$, then $\mathbf{A}$ can be seen as an adjacency matrix generated by the random dot product graph with latent position matrix $\mathbf{X} = \mathbf{Z}\mathbf{V}$, that is, $\mathbf{A} \sim \text{RDPG}(\mathbf{Z}\mathbf{V})$.

Motivated by the low-rank structure of random dot product graphs, Sussman et al. (2012) proposed to estimate the latent position matrix $\mathbf{X}$ by solving the least squares problem $\widetilde{\mathbf{X}} = \arg\min_{\mathbf{X} \in \mathbb{R}^{n \times d}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\|_{\mathrm{F}}^2$. The interpretation is that $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}$ can be viewed as the projection of the data matrix $\mathbf{A}$ onto the space of all $n \times n$ rank-d positive semidefinite matrices with regard to the Frobenius norm distance. The solution $\widetilde{\mathbf{X}}$ is referred to as the adjacency spectral embedding of $\mathbf{A}$ into $\mathbb{R}^d$, and can be computed as the matrix of eigenvectors associated with the top d eigenvalues of $\mathbf{A}$, scaled by the square roots of the corresponding eigenvalues (Eckart and Young, 1936). The asymptotic properties of $\widetilde{\mathbf{X}}$ have been established in the literature (Sussman et al., 2014; Athreya et al., 2016; Tang and Priebe, 2018). Notably, Athreya et al. (2016), Tang and Priebe (2018), and Xie and Xu (2023) have shown that each row of the adjacency spectral embedding converges to a mean-zero multivariate normal distribution after appropriate standardization.

2.2 The Surrogate Likelihood Function

In this subsection, we derive the surrogate likelihood function for the random dot product graph model. The motivation is that the exact likelihood function has a complicated structure, bringing challenges for developing the theory of maximum likelihood estimation. The difficulty partially comes from the fact that the random dot product graph model belongs to a curved exponential family, and the theory of the maximum likelihood estimation is much more difficult in curved exponential families than in canonical ones (see, for example, Section 2.3 in Bickel and Doksum, 2007). Also, the boundary of the parameter space renders the maximum likelihood estimation intractable, both computationally and analytically. To be more specific, consider the log-likelihood function

$$\ell_{\mathbf{A}}(\mathbf{X}) = \sum_{1 \leq i \leq j \leq n} \{A_{ij} \log(\rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j) + (1 - A_{ij}) \log(\rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j)\}.$$

The parameter space is defined by $\{\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d} : 0 < \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j < 1 \text{ for all } i, j\}$, and over the boundary, the log-likelihood function has an unbounded gradient. This is in sharp contrast with the requirement in Li et al. (2023), where the log-likelihood functions are required to have bounded derivatives up to the second order over the entire parameter space. These challenges motivate the development of a more computationally and analytically tractable surrogate likelihood approach.

To distinguish a generic latent position $\mathbf{x}_i \in \mathbb{R}^d$ and its true value associated with the data generating distribution, let $\mathbf{x}_{0i}$ denote the ground truth of $\mathbf{x}_i$, $i \in [n]$, and $\mathbf{X}_0 =$

$[\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}}$. Let us begin by considering the log-likelihood function of a single $\mathbf{x}_i$ when the remaining latent positions $(\mathbf{x}_{0j})_{j \neq i}$ are accessible:

$$\ell_{0in}(\mathbf{x}_i) = \sum_{j \neq i}^{n} \{A_{ij} \log(\rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j}) + (1 - A_{ij}) \log(1 - \rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j})\} \tag{1}$$
$$+ \{A_{ii} \log(\rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i) + (1 - A_{ij}) \log(1 - \rho_n \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i)\}.$$

We refer to $\ell_{0in}(\mathbf{x}_i)$ in (1) as the oracle log-likelihood function because it requires the true values of the remaining $\mathbf{x}_j$'s with $j \neq i$. Theorem 2 in Xie and Xu (2023) established the consistency and asymptotic normality of the maximizer of the oracle log-likelihood function $\ell_{0in}(\mathbf{x}_i)$ in (1). Nevertheless, the oracle log-likelihood is not computable because $(\mathbf{x}_{0j})$ are not accessible in practice. Following the idea in Xie and Xu (2023), we consider replacing the unknown latent positions by the corresponding rows of the adjacency spectral embedding. Formally, let $\widetilde{\mathbf{x}}_j$ be the jth row of the adjacency spectral embedding $\widetilde{\mathbf{X}}$, $j \in [n]$. Then we obtain the following approximation to the oracle log-likelihood:

$$\ell_{0in}(\mathbf{x}_i) \approx \sum_{j=1}^{n} \{A_{ij} \log(\rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j) + (1 - A_{ij}) \log(1 - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)\}. \tag{2}$$

Note that the last term in ℓ_{0in} is replaced by $A_{ii} \log(\rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_i) + (1 - A_{ii}) \log(1 - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_i)$ for convenience, which is immaterial. This approximation step is motivated by the uniform consistency of the adjacency spectral embedding: There exists a $d \times d$ orthogonal $\mathbf{W}$ such that $\|\widetilde{\mathbf{X}} \mathbf{W} - \rho_n^{1/2} \mathbf{X}_0\|_2 = O\{\sqrt{(\log n)/n}\}$ with high probability (Lyzinski et al., 2014; Xie, 2024).

With the approximation in (2), the constraints for the latent position $\mathbf{x}_i$ become a system of linear inequalities: $0 < \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j < 1$ for all $j \in [n]$. Geometrically, these constraints correspond to a convex polyhedron. Namely, given any vector $\mathbf{x}_i \in \mathbb{R}^d$, checking whether $\mathbf{x}_i$ is in such a convex polyhedron requires $O(n)$ operations, so that the relevant computation could be cumbersome. We now resolve this issue by applying a quadratic Taylor approximation to the terms $\log(\rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)$ and relax the parameter space for $\mathbf{x}_i$. Here we can drop the sparsity factor ρ_n without loss of generality. Formally, write $g_{ij}(\mathbf{x}_i) = A_{ij} \log(\mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)$. Then a quadratic Taylor approximation to g_{ij} at $\mathbf{x}_i = \widetilde{\mathbf{x}}_i$ leads to

$$g_{ij}(\mathbf{x}_i) = g_{ij}(\widetilde{\mathbf{x}}_i) + \frac{A_{ij} \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)}{\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j} - \frac{A_{ij} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)}{2(\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)^2} + \text{remainder}. \tag{3}$$

Meanwhile, it is also conceivable that

$$\sum_{j=1}^{n} \frac{A_{ij}}{2(\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)^2} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i) \approx \sum_{j=1}^{n} \frac{1}{2\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i) \tag{4}$$

because $\mathbb{E}_0(A_{ij}) = \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j} \approx \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j$. Hence, ignoring the constant terms that are free of $\mathbf{x}_i$, combining the approximations in (2), (3), and (4) leads to the following surrogate log-likelihood function

$$\widetilde{\ell}_{in}(\mathbf{x}_i) = \sum_{j=1}^{n} \left\{ \frac{A_{ij} \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{x}_i}{\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j} + \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{x}_i - \frac{1}{2\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j} \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{x}_i + (1 - A_{ij}) \log(1 - \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j) \right\}. \tag{5}$$

Therefore, by Cauchy–Schwarz inequality, the the parameter space for $\mathbf{x}_i$ associated with the surrogate likelihood of vertex i can be taken as the unit ball $\{\mathbf{x}_i \in \mathbb{R}^d : \|\mathbf{x}_i\| \leq 1\}$ for all $i \in [n]$ when $\max_{j \in [n]} \|\widetilde{\mathbf{x}}_j\|_2 < 1$ (which holds with high probability). Consequently, we relax the original complicated parameter space $\{\mathbf{x}_i \in \mathbb{R}^d : 0 < \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j < 1, j \in [n]\}$ to a simple unit ball, which is much more tractable to work with. Moreover, the surrogate likelihood function has a separable structure because $\widetilde{\ell}_{in}(\mathbf{x}_i)$ does not involve $\mathbf{x}_j$ for $j \neq i$. This convenience enables parallelization when related computation is requested. In addition, a simple algebra shows that the surrogate likelihood function is log-concave, a highly desired feature when optimization and Monte Carlo sampling are needed.

2.3 Comparison With the One-Step Estimator

Recently, Xie and Xu (2023) proposed a one-step estimator $\widehat{\mathbf{X}}^{\mathrm{OS}} = [\widehat{\mathbf{x}}_1^{\mathrm{OS}}, \ldots, \widehat{\mathbf{x}}_n^{\mathrm{OS}}]^{\mathrm{T}}$ for random dot product graphs that improves upon the adjacency spectral embedding:

$$\widehat{\mathbf{x}}_i^{\mathrm{OS}} = \widetilde{\mathbf{x}}_i + \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{\widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}}}{\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j (1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)} \right\}^{-1} \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{(A_{ij} - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j) \widetilde{\mathbf{x}}_j}{\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j (1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)} \right\}, \quad i \in [n]. \quad (6)$$

The one-step estimator originates from a one-step updating scheme of the Newton-Raphson method for maximizing the log-likelihood function with the initial guess being the adjacency spectral embedding (see, for example, Section 5.7 in Van der Vaart, 2000). It is clear from the construction that the one-step estimator takes advantage of the likelihood information of the sampling distribution through the Fisher information matrix and the score function.

In Section 2.2, we have shown the derivation of the surrogate log-likelihood function by applying a quadratic Taylor approximation to the logarithm function $\log(\mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)$. We now show that the same approximation treatment applied to the entire function in (2) results in an approximate log-likelihood function whose maximizer is exactly the one-step estimator. Formally, applying a second-order Taylor expansion to the term $\log(1 - \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)$ at $\mathbf{x}_i = \widetilde{\mathbf{x}}_i$ yields

$$\log \frac{(1 - \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)}{(1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)} = -\frac{\widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)}{1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j} - \frac{(\mathbf{x}_i - \widetilde{\mathbf{x}}_i)^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)}{2(1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)^2} + \text{remainder}. \quad (7)$$

Following the idea in (4), we can also conceive the following approximation:

$$\sum_{j=1}^{n} \frac{(1 - A_{ij})(\mathbf{x}_i - \widetilde{\mathbf{x}}_i)^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)}{2(1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)^2} \approx \sum_{j=1}^{n} \frac{(\mathbf{x}_i - \widetilde{\mathbf{x}}_i)^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)}{2(1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)}. \quad (8)$$

We thus obtain the following quadratic approximation to (2) modulus a constant term from (5), (7), and (8):

$$\widetilde{\ell}_{in}^{(\mathrm{OS})}(\mathbf{x}_i) = \sum_{j=1}^{n} \frac{(A_{ij} - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j) \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)}{\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j (1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)} - \sum_{j=1}^{n} \frac{(\mathbf{x}_i - \widetilde{\mathbf{x}}_i)^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{x}_i - \widetilde{\mathbf{x}}_i)}{2\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j (1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)}. \quad (9)$$

Then a simple algebra shows that the one-step estimator $\widehat{\mathbf{x}}_i^{(\mathrm{OS})}$ maximizes $\widetilde{\ell}_{in}^{(\mathrm{OS})}$ defined in (9).

Clearly, the surrogate log-likelihood function in (5) is constructed by applying a Taylor expansion to the term $\log(\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)$, whereas the one-step estimator is obtained by applying the Taylor expansion to the entire function. Thus, intuitively, the surrogate likelihood retains more likelihood information than the one-step procedure does. Below, we visualize this heuristic using a toy numerical example.

Example 2 *Consider the following random dot product graph model. Let $n = 300$, $(t_i)_{i=1}^n$ be equidistant points over $[0,1]$, $x_{0i} = 0.2 + 0.6\sin(\pi t_i)$, $i \in [n]$, and $\mathbf{X}_0 = [x_{01}, \ldots, x_{0n}]^{\mathrm{T}} \in \mathbb{R}^{n \times 1}$. Suppose $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$ and we focus on the likelihood function for $\mathbf{x}_i$ with $i = 100$. Figure 1 visualizes the comparison among the oracle log-likelihood $\ell_{0in}(\mathbf{x}_i)$, the surrogate log-likelihood $\widetilde{\ell}_{in}(\mathbf{x}_i)$, and the approximate log-likelihood $\widetilde{\ell}_{in}^{(\mathrm{OS})}(\mathbf{x}_i)$ associated with the one-step estimator. The constant terms of these functions have been added to make them comparable.*
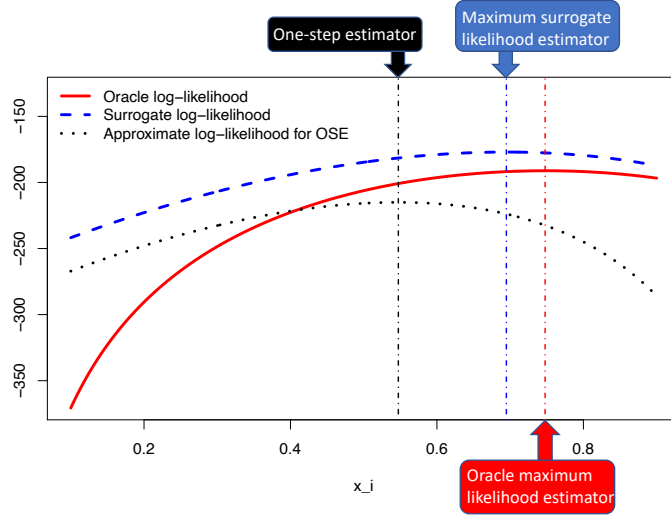

Figure 1: Comparison among the oracle log-likelihood function $\ell_{0in}(\mathbf{x}_i)$, the surrogate log-likelihood function $\widetilde{\ell}_{in}(\mathbf{x}_i)$, and the approximate log-likelihood function $\widetilde{\ell}_{in}^{(\mathrm{OS})}(\mathbf{x}_i)$ associated with the one-step estimator. The three vertical lines from left to right mark the one-step estimate, the maximum surrogate likelihood estimate, and the oracle maximum likelihood estimate, respectively.

The vertical lines mark the maximizers of the three functions, respectively. It is visually clear that the maximizer of the surrogate log-likelihood estimate is closer to that of the oracle log-likelihood than the one-step estimate is, suggesting that the maximum surrogate likelihood estimator may outperform the one-step estimator in some practical finite sample problems.

3. Maximum Surrogate Likelihood Estimation

In this section, we present the theory and the computation of maximum surrogate likelihood estimation.

3.1 Theoretical Properties

This subsection elaborates on the theoretical properties of the frequentist inference with the surrogate likelihood. Below, Theorem 3 establishes the existence and uniqueness of the maximum surrogate likelihood estimator.

Theorem 3 *Suppose* $\mathbf{A} \sim \mathrm{RDPG}(\rho_n^{1/2}\mathbf{X}_0)$ *and* $(\log n)/(n\rho_n) \to 0$ *as* $n \to \infty$. *Assume* $\lambda_d(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0/n) \geq \lambda$ *for some constant* $\lambda > 0$ *for all* $n > d$, *and* $\min_{i,j\in[n]}(\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}, 1-\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}) \geq \delta$ *for some constant* $\delta > 0$. *Let* $i \in [n]$ *be a fixed vertex and consider the maximum surrogate likelihood estimator* $\widehat{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i:\|\mathbf{x}_i\|_2\leq 1} \widetilde{\ell}_{in}(\mathbf{x}_i)$. *Then for all* $c > 0$, *there exists a constant* $N_{c,\delta,\lambda} \in \mathbb{N}_+$ *depending on* c,δ,λ *such that* $\mathbb{P}_0(\widehat{\mathbf{x}}_i$ *exists and is unique*$) \geq 1 - n^{-c}$ *for all* $n \geq N_{c,\delta,\lambda}$.

Let $\mathbf{G}_{0in} = (1/n)\sum_{j=1}^n \mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\}^{-1}$ be the Fisher information matrix with regard to the latent position $\mathbf{x}_i$. Theorem 4 below, which is one of the main results in this article, establishes the large sample properties of the maximum surrogate likelihood estimator.

Theorem 4 *Suppose the conditions of Theorem 3 hold and the embedding dimension* d *is fixed. For each* $i \in [n]$, *let* $\widehat{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i:\|\mathbf{x}_i\|_2\leq 1} \widetilde{\ell}_{in}(\mathbf{x}_i)$ *be the maximum surrogate likelihood estimator. Then there exists an orthogonal matrix* $\mathbf{W} \in \mathbb{O}(d)$ *that depends on* n, *such that for each* $i \in [n]$,

$$\sqrt{n}\mathbf{G}_{0in}^{1/2}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d).$$

Furthermore, if $(\log n)^4/(n\rho_n) \to 0$, *then*

$$\|\widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 - \frac{1}{n}\sum_{i=1}^n \mathrm{tr}(\mathbf{G}_{0in}^{-1}) \to 0$$

in probability, where $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1,\ldots,\widehat{\mathbf{x}}_n]^{\mathrm{T}}$.

Remark 5 (Sparsity condition) *The sparsity condition that* $(\log n)/(n\rho_n) \to 0$ *required in Theorem 3 and in the asymptotic normality of Theorem 4 is minimal in the following sense. It is well known that the random adjacency matrix* $\mathbf{A}$ *no longer concentrates around its expected value* $\mathbb{E}_0(\mathbf{A})$ *when* $(\log n)/(n\rho_n) \to \infty$ *(Tang and Priebe, 2018). Furthermore, Abbe et al. (2020) and Xie (2024) showed that in order to have* $\|\rho_n^{-1/2}\widetilde{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{2\to\infty} = o(1)$ *with high probability, which is an indispensable ingredient in our employed proof, it is necessary that* $(\log n)/(n\rho_n) \to 0$.

Remark 6 (Comparison with other estimators) *Athreya et al. (2016), Tang and Priebe (2018), and Xie and Xu (2023) have establish the large sample properties of the adjacency spectral embedding and the one-step estimator as the following. Let* $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1,\ldots,\widetilde{\mathbf{x}}_n]^{\mathrm{T}}$ *and* $\widehat{\mathbf{X}}^{(\mathrm{OS})} = [\widehat{\mathbf{x}}_1^{(\mathrm{OS})},\ldots,\widehat{\mathbf{x}}_n^{(\mathrm{OS})}]^{\mathrm{T}}$. *Under appropriate conditions, for each vertex* $i \in [n]$,

$$\sqrt{n}\boldsymbol{\Sigma}_{in}^{-1/2}(\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d), \quad \sqrt{n}\mathbf{G}_{0in}^{1/2}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d),$$

$$\|\widetilde{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 - \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}(\boldsymbol{\Sigma}_{in}) \overset{\mathbb{P}_0}{\to} 0, \quad \|\widehat{\mathbf{X}}^{(\mathrm{OS})}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 - \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}(\mathbf{G}_{0in}^{-1}) \overset{\mathbb{P}_0}{\to} 0.$$

where the covariance matrix $\boldsymbol{\Sigma}_{in}$ satisfies $\boldsymbol{\Sigma}_{in} \succeq \mathbf{G}_{0in}^{-1}$. Theorem 4 thus suggests that the maximum surrogate likelihood estimator improves upon the adjacency spectral embedding and is (first-order) asymptotically equivalent to the one-step estimator. This phenomenon is also known as the local efficiency (Xie and Xu, 2023) because the asymptotic covariance matrix for a single latent position $\mathbf{x}_i$ is the same as that of the oracle maximum likelihood estimator.

Given that both the one-step estimator and the maximum surrogate likelihood estimator achieve the local efficiency, the comparison at the first-order ($O(n^{-1/2})$) is unable to distinguish their performance. To further discern the difference between these two estimators, it is desirable to explore their second-order ($O(n^{-1})$) behavior. Such an idea can be formalized by the second-order ($O(n^{-1})$) bias of an estimator. Generically, given an asymptotically unbiased estimator $\widehat{\boldsymbol{\theta}}_n$ for an unknown parameter $\boldsymbol{\theta}$ (that is, $\lim_{n\to\infty}\mathbb{E}\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$), if $\mathbb{E}\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta} + \mathbf{b}_n + o(n^{-1})$ and $\mathbf{b}_n = O(n^{-1})$, then the $O(n^{-1})$ bias of $\widehat{\boldsymbol{\theta}}_n$ is given by $\mathrm{Bias}(\widehat{\boldsymbol{\theta}}_n) = \mathbf{b}_n$. Also see Pfanzagl and Wefelmeyer (1978); Rilstone et al. (1996); Newey and Smith (2004); Schennach (2007) for the analyses of the $O(n^{-1})$ biases in the econometric literature. Below, Theorem 7 establishes the formulae of the $O(n^{-1})$ biases of the one-step estimator and the maximum surrogate likelihood estimator.

Theorem 7 *Suppose the conditions of Theorem 3 hold and the embedding dimension d is fixed. Further assume that $\rho_n = 1$ and $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$ is a diagonal matrix with distinct eigenvalues, and the differences of these are lower bounded by a constant multiple of n. For each $i \in [n]$, let $\widehat{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i:\|\mathbf{x}_i\|_2\leq 1} \widetilde{\ell}_{in}(\mathbf{x}_i)$ be the maximum surrogate likelihood estimator. Then the $O(n^{-1})$ biases of $\widehat{\mathbf{x}}_i$ and $\widehat{\mathbf{x}}_i^{(\mathrm{OS})}$ are given by*

$$\mathrm{Bias}(\widehat{\mathbf{x}}_i) = \mathbf{b}_i^{(\mathrm{MSLE})} + \mathbf{b}_i^{(\mathrm{ASE})} + \mathbf{b}_i^{(\mathrm{base})},$$
$$\mathrm{Bias}(\widehat{\mathbf{x}}_i^{(\mathrm{OS})}) = \mathbf{b}_i^{(\mathrm{OS})} + \mathbf{b}_i^{(\mathrm{ASE})} + \mathbf{b}_i^{(\mathrm{base})},$$

where

$$\mathbf{b}_i^{(\mathrm{OSE})} = \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}\frac{(2\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j} - 1)\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{0j}}{(\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})(1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})},$$

$$\mathbf{b}_i^{(\mathrm{MSLE})} = -\mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}\frac{(1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}} + \mathbf{G}_{0in}^{-1}\frac{1}{n^2}\sum_{j=1}^{n}\frac{\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j}}{(1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})^2},$$

$$\mathbf{b}_i^{(\mathrm{base})} = -\mathbf{G}_{0in}^{-1}\frac{1}{n^2}\sum_{j=1}^{n}\left\{\frac{-1}{(\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})^2} + \frac{1}{(1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})^2}\right\}\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j},$$

$$\mathbf{b}_i^{(\mathrm{ASE})} = -\mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{\mathbf{x}_{0j}x_{0ik}\beta_{jk}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}$$
$$+ \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}\frac{2\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j} - 1}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}\mathbf{x}_{0i}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{0i} + \mathbf{G}_{in}^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{0i}$$

10

$$- \mathbf{G}_{0in}^{-1} \frac{1}{n^2} \sum_{j=1}^{n} \frac{(2\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j} - 1)\mathbf{x}_{0j}\mathbf{x}_{0i}^{\mathrm{T}}\boldsymbol{\Sigma}_{jn}\mathbf{x}_{0i}}{(\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})^2(1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})^2} - \mathbf{G}_{0in}^{-1} \frac{1}{n^2} \sum_{j=1}^{n} \frac{\boldsymbol{\Sigma}_{jn}\mathbf{x}_{0i}}{(\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})^2(1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})^2},$$

$$\beta_{jk} = \mathbf{e}_j^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{u}_k\mathbf{u}_k^{\mathrm{T}} - \sum_{m \in [d]\backslash\{k\}} \frac{\lambda_m \mathbf{u}_m \mathbf{u}_m^{\mathrm{T}}}{\lambda_k - \lambda_m}\right)\mathrm{diag}\left\{\left(\sum_{b=1}^{n} \frac{\mathbf{x}_{0a}^{\mathrm{T}}\mathbf{x}_{0b}(1 - \mathbf{x}_{0a}^{\mathrm{T}}\mathbf{x}_{0b})x_{0bk}}{\lambda_k^2}\right)_{a=1}^{n}\right\},$$

and $\mathbf{u}_k$ is the eigenvector of $\mathbf{P}_0 = \mathbf{X}_0\mathbf{X}_0^{\mathrm{T}}$ associated with the kth largest eigenvalue λ_k.

Remark 8 (Interpretation of the bias terms) *Each term in the $O(n^{-1})$ bias formula has an interesting interpretation. The term $[\beta_{j1}, \ldots, \beta_{jd}]^{\mathrm{T}}$ is precisely the $O(n^{-1})$ bias of the ASE and has been obtained in Xie and Zhang (2025). The term $\mathbf{b}_i^{(\mathrm{ASE})}$ stems from the substitution of the unknown $\mathbf{x}_{0j}$'s with the ASE $\widetilde{\mathbf{x}}_j$'s and naturally connects to the $O(n^{-1})$ bias of the ASE. This term does not depend on whether the one-step estimator or the maximum surrogate likelihood estimator is used. It also vanishes in the hypothetical scenario where the oracle knowledge of $\mathbf{x}_{0j}$'s is accessible. The term $\mathbf{b}_i^{(\mathrm{base})}$ is intrinsic to the use of maximum likelihood principle and does not vanish even if the oracle maximum likelihood estimator (that is, $\arg\max_{\mathbf{x}_i} \ell_{0in}(\mathbf{x}_i)$) is used. The key difference of the $O(n^{-1})$ biases of these estimators lies in $\mathbf{b}_i^{(\mathrm{MSLE})}$ and $\mathbf{b}_i^{(\mathrm{OS})}$, which explains how the former estimator retains more likelihood information than the latter estimator.*

Example 3 *Consider the following rank-one two-block stochastic block model with block probability matrix*

$$\mathbf{B} = \begin{bmatrix} p^2 & pq \\ pq & q^2 \end{bmatrix}$$

and cluster assignment function $\tau(\cdot)$ defined by $\tau(i) = 1$ if $i \in \{1, \ldots, n/2\}$ and $\tau(i) = 2$ if $i \in \{n/2 + 1, \ldots, n\}$ with $n = 300$. For this specific model, we compute the ratio of sum of squared biases $\sum_{i=1}^{n} \mathrm{Bias}^2(\widehat{\mathbf{x}}_i^{(\mathrm{OS})})/\sum_{i=1}^{n} \mathrm{Bias}^2(\widehat{\mathbf{x}}_i)$ as a function of p and q, where p, q vary over $[0.05, 0.95]$ and visualize the ratio in 2. The plot shows that the maximum surrogate likelihood estimator results in less bias compared to the one-step estimator for a broad range of (p, q) values.

3.2 Computation Details

This subsection discusses the detailed algorithm for computing the maximum surrogate likelihood estimator. For a given vertex $i \in [n]$, the estimator $\widehat{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i} \widetilde{\ell}_{in}(\mathbf{x}_i)$ can be computed separately for each vertex $i \in [n]$. Thus, it is sufficient to design an algorithm for solving the optimization problem

$$\max_{\|\mathbf{x}_i\|_2 \leq 1} \frac{1}{n}\widetilde{\ell}_{in}(\mathbf{x}_i). \tag{10}$$

Then the entire estimator $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$ for all vertices can be obtained through a parallelization over $i \in [n]$. Let us consider the optimization problem (10). Observe that the objective function $(1/n)\widetilde{\ell}_{in}(\mathbf{x}_i)$ is concave and can be written in a sample average fashion,
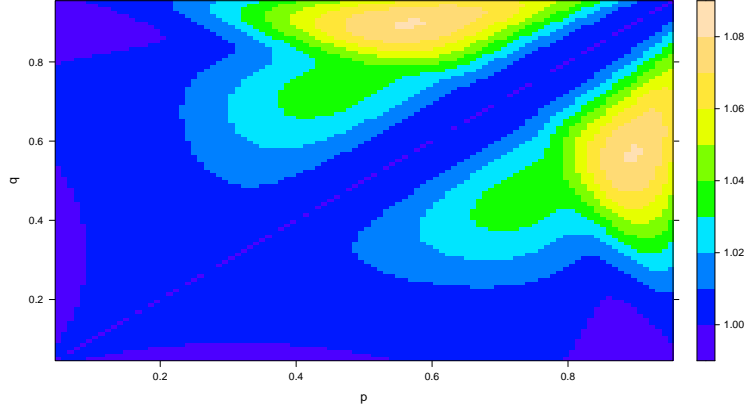
Figure 2: Level plot of the ratio of sum of squared biases $\sum_{i=1}^{n} \text{Bias}^2(\widehat{\mathbf{x}}_i^{(\text{OS})}) / \sum_{i=1}^{n} \text{Bias}^2(\widehat{\mathbf{x}}_i)$ as a function of p and q, where p, q vary over $[0.05, 0.95]$ for Example 3.

which motivates us to adopt the stochastic gradient descent algorithm (Robbins and Monro, 1951). Let $j_1, \ldots, j_s$ be independent $\text{Unif}(1, \ldots, n)$ random variables, where $s \in \{1, \ldots, n\}$ is the so-called batch size, and for any $j \in [n]$, let

$$m_i(\mathbf{x}_i, j) = \frac{A_{ij}\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{x}_i}{\widetilde{p}_{ij}} + \widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{x}_i - \frac{1}{2\widetilde{p}_{ij}}\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{x}_i + (1 - A_{ij})\log(1 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j).$$

It is clear that for each j_k, $k \in [s]$, $m_i(\mathbf{x}_i, j_k)$ can be viewed as a noisy measurement of the objective function $(1/n)\widetilde{\ell}_{in}(\mathbf{x}_i)$ because $(1/n)\widetilde{\ell}_{in}(\mathbf{x}_i) = \mathbb{E}_{j_k}\{m_i(\mathbf{x}_i, j_k)\}$. Then given a sequence of step sizes $\{\alpha_t\}_{t \geq 1}$ and a initial guess $\widehat{\mathbf{x}}_i^{(0)}$, the stochastic gradient descent algorithm generates a sequence of iterates $\{\widehat{\mathbf{x}}_i^{(t)}\}_{t \geq 1}$ using the updating scheme

$$\widehat{\mathbf{x}}_i^{(t+1)} = \widehat{\mathbf{x}}_i^{(t)} + \frac{\alpha_t}{s}\sum_{k=1}^{s}\frac{\partial m_i}{\partial \mathbf{x}}(\widehat{\mathbf{x}}_i^{(t)}, j_k^{(t)}), \tag{11}$$

where $\{(j_1^{(t)}, \ldots, j_s^{(t)})\}_{t \geq 1}$ are independent copies of $(j_1, \ldots, j_s)$. The advantage of the stochastic gradient descent method over the classical gradient descent algorithm is that, with a comparatively small batch size s, one only needs to compute s gradient measurements of $m_i(\mathbf{x}_i, j)$ rather than all the gradient measurements of $\{m_i(\mathbf{x}_i, j)\}_{j=1}^{n}$. This computational convenience is especially desired when the network contains large number of vertices. To implement the algorithm with adaptive step sizes, we follow the suggestion given by Duchi et al. (2011) and Li and Orabona (2019) and take

$$\alpha_t = a_0 \left\{ b_0 + \sum_{l=1}^{t-1}\left\|\frac{1}{s}\sum_{k=1}^{s}\frac{\partial m_i}{\partial \mathbf{x}}(\widehat{\mathbf{x}}_i^{(l)}, j_k^{(l)})\right\|_2^2 \right\}^{-(\epsilon+1/2)}, \tag{12}$$

where $a_0, b_0 > 0$ and $0 < \epsilon \leq 1/2$ are constants.

The key difference between our algorithm and the standard stochastic gradient descent algorithm is that the feasible region $\{\mathbf{x}_i \in \mathbb{R}^d : \|\mathbf{x}_i\| \leq 1\}$ is compact. Therefore, whenever an updated value $\widehat{\mathbf{x}}_i^{(t+1)}$ stays outside the feasible region, one repeats step-halving procedures until $\|\widehat{\mathbf{x}}_i^{(t+1)}\| \leq 1$. We present the detailed stochastic gradient descent algorithm for computing the maximum surrogate likelihood estimator in Algorithm 1, the convergence of which is guaranteed by Theorem 9 below.

Algorithm 1 Stochastic gradient descent for maximum surrogate likelihood estimation

---

1: **Input:** The adjacency matrix $\mathbf{A} = [A_{ij}]_{n \times n}$ and the embedding dimension d.
2: **Set:** Tuning parameters $a_0, b_0 > 0$, $\epsilon \in (0, 1/2]$, and batch size $1 \leq s \leq n$.
3: Compute the spectral decomposition of the adjacency matrix $\mathbf{A} = \sum_{i=1}^{n} \widehat{\lambda}_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_j^{\mathrm{T}}$, where $|\widehat{\lambda}_1| \geq |\widehat{\lambda}_2| \geq \ldots \geq |\widehat{\lambda}_n|$, and $\widehat{\mathbf{u}}_i^{\mathrm{T}} \widehat{\mathbf{u}}_j = \mathbb{1}(i = j)$ for all $i, j \in [n]$.
4: Compute the adjacency spectral embedding:

$$\widetilde{\mathbf{X}} = \widehat{\mathbf{X}}^{\mathrm{ASE}} = [\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_d] \cdot \mathrm{diag}(|\widehat{\lambda}_1|^{1/2}, \ldots, |\widehat{\lambda}_d|^{1/2}),$$

and write $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d}$. Let $= \widetilde{p}_{ij} = \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j$ for all $i, j \in [n]$.
5: For $i = 1, 2, \ldots, n$
6:     Initialize $\widehat{\mathbf{x}}_i^{(1)} = \widetilde{\mathbf{x}}_i$.
7:     Set the iteration counter $t = 1$.
8:     While not converge
9:         Sample without replacement $j_1, j_2, \ldots, j_s \sim \mathrm{Unif}(1, 2, \ldots, n)$.
10:         Compute the average gradient at $\widehat{\mathbf{x}}_i^{(t)}$

$$\bar{\mathbf{g}}^{(t)}(\widehat{\mathbf{x}}_i^{(t)}) = \frac{1}{s} \sum_{k=1}^{s} \frac{\partial m_i}{\partial \mathbf{x}_i}(\mathbf{x}_i, j_k) \bigg|_{\mathbf{x}_i = \widehat{\mathbf{x}}_i^{(t)}}.$$

11:         Compute the step size α_t using formula (12).
12:         Compute $\widehat{\mathbf{x}}_i^{(t+1)} = \widehat{\mathbf{x}}_i^{(t)} + \alpha_t \bar{\mathbf{g}}^{(t)}(\widehat{\mathbf{x}}_i^{(t)})$.
13:         If $\|\widehat{\mathbf{x}}_i^{(t+1)}\|_2 > 1$, then set $\alpha_t \longleftarrow \alpha_t/2$ and go to line 12
14:         Set $t \longleftarrow t + 1$.
15:     End While
16: End For
17: **Output:**   The MSLE $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n]^{\mathrm{T}}$.

Theorem 9 *Let the vertex $i \in [n]$ be fixed and suppose $(1/n)\widetilde{\ell}_{in}(\mathbf{x}_i)$ is well-defined. Assume that $\widehat{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i : \|\mathbf{x}_i\| \leq 1} (1/n)\widetilde{\ell}_{in}(\mathbf{x}_i)$ lies in the interior of $\{\mathbf{x}_i \in \mathbb{R}^d : \|\mathbf{x}_i\| \leq 1\}$. Then the sequence of iterates $\{\widehat{\mathbf{x}}_i^{(t)}\}_{t \geq 1}$ generated by (11) with step sizes $\{\alpha_t\}_{t \geq 1}$ given by (12) and step-halving converges to $\widehat{\mathbf{x}}_i$ almost surely with regard to the distribution of $\{(j_1^{(t)}, \ldots, j_s^{(t)})\}_{t \geq 1}$.*

Remark 10 *The surrogate log-likelihood function $\widetilde{\ell}_{in}(\mathbf{x}_i)$ is well-defined only when $\mathbf{x}^{\mathrm{T}}\widetilde{\mathbf{x}}_j < 1$ for all $j \in [n]$ because of the logarithm terms $\{\log(1 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)\}_{j=1}^{n}$. For sufficiently large*

n, the constraint is satisfied by requiring that $\|\mathbf{x}_i\|_2 \leq 1$ since the adjacency spectral embedding $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}}$ satisfies $\max_{j\in[n]} \|\widetilde{\mathbf{x}}_j\|_2 < 1$ with high probability. However, this requirement may not hold in certain finite sample problems, in which case the surrogate log-likelihood function $\widetilde{\ell}_{in}(\mathbf{x}_i)$ is no longer well-defined. This numerical issue can be practically addressed by the following smooth concatenation technique. Roughly speaking, for a fixed $j \in [n]$, when $1 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j$ drops below a small threshold, we replace the objective function $(1/n)\widetilde{\ell}_{in}(\mathbf{x}_i)$ by a quadratic function such that the two pieces of functions are concatenated smoothly. Formally, let $\tau > 0$ be a small threshold and define

$$h_i(\mathbf{x}_i, j) = \begin{cases} m_i(\mathbf{x}_i, j), & \text{if } 1 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j \geq \tau, \\ \alpha_{ij}(\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)^2 + \beta_{ij}(\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j) + \gamma_{ij}, & \text{if } 1 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j < \tau, \end{cases} \tag{13}$$

for each $j \in [n]$, where $\alpha_{ij}, \beta_{ij}, \gamma_{ij}$ are coefficients such that $h_i(\cdot, j)$ is twice continuously differentiable. Then the objective function $(1/n)\widetilde{\ell}_{in}(\mathbf{x}_i)$ can be replaced by $(1/n)\sum_{j=1}^{n} h_i(\mathbf{x}_i, j)$ and the aforementioned stochastic gradient descent algorithm applies with $\partial m_i(\mathbf{x}_i, j)/\partial \mathbf{x}_i$ replaced by $\partial h_i(\mathbf{x}_i, j)/\partial \mathbf{x}_i$.

4. Bayesian Estimation With Surrogate Likelihood

This section explores Bayesian estimation of random dot product graphs with the proposed surrogate likelihood. Although Xie and Xu (2020) has established the minimax optimality of the Bayesian random dot product graph model with the exact likelihood, the asymptotic shape of the posterior distribution is yet to be characterized because of the complicated structure of the exact likelihood function. In contrast, thanks to the separable and log-concave properties of the surrogate likelihood, we are able to completely characterize the asymptotic posterior distribution of the latent positions with the exact likelihood replaced by the surrogate likelihood. Formally, for any fixed vertex $i \in [n]$ and a prior distribution $\pi(\cdot)$ supported on $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$, the posterior distribution of $\mathbf{x}_i$ given $\mathbf{A}$ with the surrogate log-likelihood function $\widetilde{\ell}_{in}(\mathbf{x}_i)$ can be written as

$$\widetilde{\pi}_{in}(\mathbf{x}_i \mid \mathbf{A}) = \frac{\exp\{\widetilde{\ell}_{in}(\mathbf{x}_i)\}\pi(\mathbf{x}_i)}{\int \exp\{\widetilde{\ell}_{in}(\mathbf{x}_i)\}\pi(\mathbf{x}_i)\mathrm{d}\mathbf{x}_i}. \tag{14}$$

Then the joint posterior density of the entire latent position matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}}$ is taken as the product $\widetilde{\pi}_n(\mathbf{X} \mid \mathbf{A}) = \prod_{i=1}^{n} \widetilde{\pi}_{in}(\mathbf{x}_i \mid \mathbf{A})$ because the surrogate log-likelihood function for the whole graph is separable across different vertices.

When the exact likelihood function is not available or intractable for analysis or computation, the idea of using a general statistical criterion function to replace the likelihood in the Bayes formula is not entirely new, among which an influential work is Chernozhukov and Hong (2003). There have also been several recent works addressing the large sample properties of the so-called quasi-posterior or Gibbs posterior distributions (Kleijn and van der Vaart, 2012; Miller, 2021; Syring and Martin, 2018, 2022). One key difference is that unlike the well-specified exact posterior distributions, the frequentist coverage of the credible sets of the quasi-posterior distributions may not agree with their credibility level (Kleijn and van der Vaart, 2012). Below, we show that, with the surrogate likelihood, the

posterior distribution produces credible sets that have the correct frequentist coverage. This is achieved through the following Bernstein–von Mises theorem.

Theorem 11 *Suppose the conditions of Theorem 3 hold and the embedding dimension d is fixed. Let $\pi(\cdot)$ be a prior density satisfying $c \leq \pi(\mathbf{x}_i) \leq C$ and $|\pi(\mathbf{x}) - \pi(\mathbf{y})| \leq C' \|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y}$ with $\|\mathbf{x}\|_2, \|\mathbf{y}\|_2 \leq 1$ for some constants $0 < c, C, C' < \infty$. Let $\mathbf{W}$ be the $d \times d$ orthogonal matrix in Theorem 4. For any fixed vertex $i \in [n]$, let $\widehat{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i : \|\mathbf{x}_i\|_2 \leq 1} \widetilde{\ell}_{in}(\mathbf{x}_i)$, $\mathbf{t} = \sqrt{n} \mathbf{W}^{\mathrm{T}}(\mathbf{x}_i - \widehat{\mathbf{x}}_i)$, and $\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A})$ be the density of $\mathbf{t}$ induced from (14). Then for all $\alpha > 0$,*

$$\max_{i \in [n]} \int (1 + \|\mathbf{t}\|_2^\alpha) \left| \widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A}) - \det(2\pi \mathbf{G}_{0in}^{-1})^{-1/2} e^{-\mathbf{t}^{\mathrm{T}} \mathbf{G}_{0in} \mathbf{t}/2} \right| \mathrm{d}\mathbf{t} \xrightarrow{\mathbb{P}_0} 0. \tag{15}$$

Below, Corollary 12 discusses the effect of Theorem 11 on subsequent inference. In particular, it shows that for each vertex $i \in [n]$, the posterior mean has the same asymptotic distribution as the maximum surrogate likelihood estimator, and the asymptotic level-α credible set has the correct frequentist coverage probability.

Corollary 12 *Suppose the conditions of Theorem 11 hold. For any $i \in [n]$, let $\mathbf{x}_i^* = \int \mathbf{x}_i \widetilde{\pi}_{in}(\mathbf{x}_i \mid \mathbf{A}) \mathrm{d}\mathbf{x}_i$ and $\mathbf{\Sigma}_{in}^* = \int (\mathbf{x}_i - \mathbf{x}_i^*)(\mathbf{x}_i - \mathbf{x}_i^*)^{\mathrm{T}} \widetilde{\pi}_{in}(\mathbf{x}_i \mid \mathbf{A}) \mathrm{d}\mathbf{x}_i$ be the posterior mean and covariance matrix of $\mathbf{x}_i$, respectively, and $\mathbf{X}^* = [\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*]^{\mathrm{T}}$. Let $q_{1-\alpha}$ be the $(1 - \alpha)$ quantile of the χ_d^2 distribution and $\mathcal{C}_{in}(\alpha) = \{\mathbf{x}_i : (\mathbf{x}_i - \mathbf{x}_i^*)^{\mathrm{T}} (\mathbf{\Sigma}_{in}^*)^{-1} (\mathbf{x}_i - \mathbf{x}_i^*) \leq q_{1-\alpha}\}$ be the asymptotic $(1 - \alpha)$-credible set for $\mathbf{x}_i$, where $\mathbf{W} \in \mathbb{O}(d)$ is given in Theorem 4. Then*

$$\sqrt{n} \mathbf{G}_{0in}^{1/2} (\mathbf{W}^{\mathrm{T}} \mathbf{x}_i^* - \rho_n^{1/2} \mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d)$$

and $\mathbb{P}_0 \{\rho_n^{1/2} \mathbf{W} \mathbf{x}_{0i} \in \mathcal{C}_{in}(\alpha)\} \to 1 - \alpha$. Furthermore, if $(\log n)^4 / (n\rho_n) \to 0$, then

$$\|\mathbf{X}^* \mathbf{W} - \rho_n^{1/2} \mathbf{X}_0\|_{\mathrm{F}}^2 - \frac{1}{n} \sum_{i=1}^n \mathrm{tr}(\mathbf{G}_{0in}^{-1}) \xrightarrow{\mathbb{P}_0} 0.$$

In practice, the posterior distribution based on the surrogate likelihood can be computed using a standard Metropolis–Hastings algorithm with parallelization over the vertices $i \in [n]$. The detailed algorithm is provided in the Supplementary Material. Note that in practice, we can also apply the smooth concatenation technique discussed in Remark 10 to the posterior computation by simply replacing the surrogate log-likelihood function $\widetilde{\ell}_{in}(\mathbf{x}_i)$ in the Bayes formula (14) by $\sum_{j=1}^n h_i(\mathbf{x}_i, j)$ defined in (13).

5. Numerical Examples

In this section, we present some simulated examples and some empirical data examples.

5.1 A Latent Curve Example

In this subsection, we study the empirical performance of the proposed estimation procedures through a simulated random dot product graph example, where the latent positions are generated from a one-dimensional curve. Consider a random dot product graph with

n vertices and latent dimension $d = 1$. For each vertex $i \in [n]$, the latent position x_{0i} for the ith vertex is set to $x_{0i} = 0.8 \sin\{\pi(i-1)/(n-1)\} + 0.1$. Let $\mathbf{X}_0 = [x_{01}, \ldots, x_{0n}]^{\mathrm{T}}$, $n = 1000$. Given $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$, we consider the following four estimation procedures for $\mathbf{X}_0$: the adjacency spectral embedding (ASE), the one-step estimate (OSE), the maximum surrogate likelihood estimate (MSLE) obtained using the step-halving stochastic gradient descent algorithm, and the Bayes estimate with the surrogate likelihood (BE). For the Bayes estimate, we use the uniform prior on the unit disk for all $\mathbf{x}_i$. The Metropolis–Hastings sampler is implemented with parallelization over vertices $i \in [n]$, and each Markov chain contains 1000 burn-in iterations and 2000 post-burn-in samples with a thinning of 5. The posterior mean is taken as the point estimate. The convergence diagnostics of the Markov chains are provided in the Supplementary Material, showing no signs of non-convergence.

The performance of the above estimates is investigated via the following two objectives: The recovery of the latent position matrix $\mathbf{X}_0$; The empirical coverage probabilities of the vertex-wise confidence intervals based on the MSLE and credible intervals based on the BE. Specifically, for the first objective, given a generic estimate $\bar{\mathbf{X}}$ for $\mathbf{X}_0$, we use the sum of squared errors (SSEs) $\inf_{\mathbf{W} \in \{\pm 1\}} \|\bar{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$ as the evaluation metric. For the second objective, we compute the vertex-wise asymptotic 95% frequentist confidence intervals and Bayesian credible intervals. The vertex-wise 95% confidence intervals based on the MSLE are computed as follows: Denote the $1 - \alpha/2$ quantile of the standard normal distribution by $z_{1-\alpha/2}$. Then by Theorem 4, for each $i \in [n]$, the $(1 - \alpha)$ confidence interval for x_{0i} is $(|\widehat{x}_i| - \{n\widehat{\mathbf{G}}(\widehat{x}_i)\}^{-1/2} z_{1-\alpha/2}, |\widehat{x}_i| + \{n\widehat{\mathbf{G}}(\widehat{x}_i)\}^{-1/2} z_{1-\alpha/2})$, where $\widehat{\mathbf{G}}_{in}(\widehat{x}_i) = (1/n) \sum_{j=1}^{n} \widehat{x}_j \{\widehat{x}_i(1 - \widehat{x}_i \widehat{x}_j)\}^{-1}$ is the plug-in estimate of the asymptotic variance. The vertex-wise 95% credible intervals based on the posterior distribution with the surrogate likelihood function can be obtained directly from the Metropolis–Hastings samples. The same numerical experiment is repeated for 1000 Monte Carlo replicates.

Estimate	ASE	OSE	MSLE	BE
SSE	0.4707	0.4592	0.4596	0.4608
Standard error of SSE	0.0216	0.0209	0.0209	0.0210

Two-sample t-test		ASE vs OSE	ASE vs MSLE	ASE vs BE
p-value		1.0×10^{-32}	7×10^{-31}	8×10^{-25}

Table 1: The average SSEs, their standard errors, and the p-values of the two-sample t-tests between the SSEs of the ASE against the remaining three estimates for Section 5.1 with $n = 1000$.

For the first objective, the SSEs of the estimates are shown in Table 1. We can see that the SSEs of the adjacency spectral embedding is comparatively larger than those of the remaining competitors, while the likelihood-based estimates have smaller SSEs. The p-values of the pairwise two-sample t-tests among the SSEs of these estimates are tabulated in Table 1 as well and they show that the differences between the ASE and the remaining likelihood-based estimates are statistically significant. This phenomenon empirically validate the conclusion that the likelihood-based estimates, namely, the OSE, the MSLE, and the BE, improve upon the the spectral-based adjacency spectral embedding.
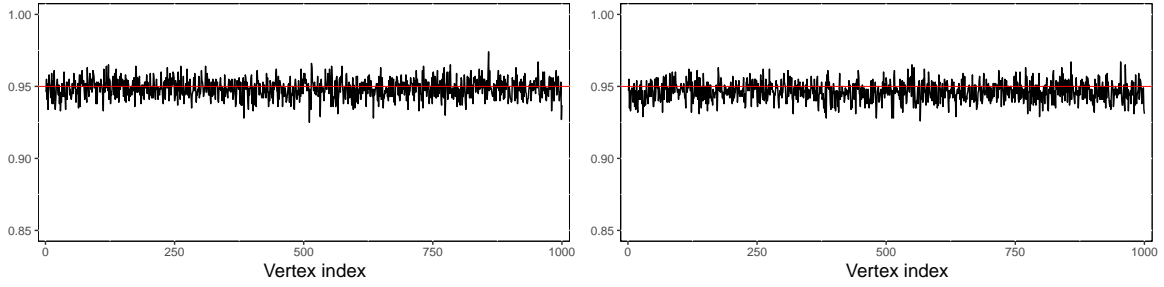
(a) Coverage probabilities of confidence intervals   (b) Coverage probabilities of credible intervals

Figure 3: Numerical results for Section 5.1: Panels (a) and (b) present the empirical coverage probabilities of the 95% confidence intervals constructed based on the MSLE and the 95% credible intervals constructed from the Metropolis–Hastings samples, respectively, where the red horizontal lines mark the 95% nominal coverage probability.

For the second objective, Figure 3 (a) and (b) visualize the empirical coverage probabilities of the vertex-wise 95% confidence intervals based on the MSLE and the vertex-wise 95% Bayesian credible intervals across the 1000 Monte Carlo replicates, respectively. It is clear that the empirical coverage probabilities of these confidence intervals and credible intervals are close to the nominal 95% level, validating the theory developed in Section 3 and Section 4.

In addition to the above investigation in a large sample regime with $n = 1000$, we also explore the performance of the proposed estimation methods in a comparatively small sample regime with $n = 30$. Here, we focus on the performance of different estimates using the SSE as the evaluation metric. Besides the aforementioned four estimates, we also consider the maximum likelihood estimate (MLE). Note that although the theory of the MLE is still open, it is always possible to find a local maximizer of the likelihood function using any optimization toolkit. Here we use the R built-in `optim` function in practice. We repeat the same numerical experiment for 1000 independent Monte Carlo replicates, visualize the boxplots of the SSEs in Figure 4, and tabulate the numeric values of the summary statistics of these SSEs in Table 2. We can see that in this small sample scenario, the MSLE and the OSE do not outperform the baseline ASE and the MLE as they have larger SSEs, while the BE has the least SSEs. The p-values of the pairwise t-test of the SSEs of the BE against those of the remaining competitors are reported in Table 2 as well, showing that the differences between BE and the other competitors are statistical significant. This observation shows the potential advantage of the Bayesian estimation procedure based on the Markov chain Monte Carlo sampling algorithm over the classical optimization-based estimation methods for finite-sample problems in practice.

5.2 A Rank-Two Random Dot Product Graph Example

We now consider a rank-two random dot product graph with $n = 300$ vertices and latent dimension $d = 2$, where the latent positions $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}}$ are given by $\mathbf{x}_{0i} = [0.15\sin\{\pi(i-1)/(n-1)\} + 0.6, 0.15\cos\{\pi(i-1)/(n-1)\} + 0.6]^{\mathrm{T}}$. Similar to Section 5.1,

Estimate	ASE	OSE	MSLE	MLE	BE
SSE	0.4594	0.4608	0.5739	0.4451	0.3886
Standard error of SSE	0.1204	0.1439	0.2409	0.1139	0.1079
Computation time (seconds)	9×10^{-3}	4.8×10^{-2}	139	5.1	148
Two-sample t-test	BE vs ASE	BE vs OSE	BE vs MSLE	BE vs MLE	
p-value	1.1×10^{-41}	1.9×10^{-35}	1.2×10^{-93}	3.7×10^{-29}	

Table 2: The average SSEs, their standard errors, and the p-values of the two-sample t-tests between the SSEs of the BE against the remaining estimates for Section 5.1 with $n = 30$,

for a given $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$, we also implement the adjacency spectral embedding (ASE), the one-step estimate (OSE), the maximum surrogate likelihood estimate (MSLE), and the Bayes estimate with the surrogate likelihood (BE). The implementation details are the same as those of 5.1, and we compute the sum of squared errors (SSEs) $\inf_{\mathbf{W} \in \mathbb{O}(d)} \|\bar{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$ as the evaluation metric. The same numerical experiment is repeated for 1000 Monte Carlo replicates. The average SSEs, their standard errors, and the computation times across repeated experiments are summarized in Table 3. The differences in SSEs are statistically

Estimate	ASE	OSE	MSLE	BE
SSE	14.94	21.70	13.34	10.91
Standard error of SSE	0.4326	8.057	0.5929	0.6229
Computation time (seconds)	0.005	0.01	4.80	41.43

Table 3: The average SSEs, their standard errors, and the computation times for Section 5.2 with $n = 300$.

significant. It is clear that the Bayes estimate with the surrogate likelihood results in the best performance in terms of SSE, while both the maximum surrogate likelihood estimate and the Bayes estimate outperform the baseline adjacency spectral embedding and the one-step estimate. Note, nonetheless, that the performance improvement of the proposed methods is at the cost of additional computation times.

5.3 A Stochastic Block Model Example

We now consider a stochastic block model in the context of a random dot product graph. The latent dimension is $d = 2$, the number of communities is $K = 5$, and the unique latent positions are $\mathbf{v}_1 = [0.3, 0.3]^{\mathrm{T}}$, $\mathbf{v}_2 = [0.5, 0.5]^{\mathrm{T}}$, $\mathbf{v}_3 = [0.7, 0.7]^{\mathrm{T}}$, $\mathbf{v}_4 = [0.3, 0.7]^{\mathrm{T}}$, and $\mathbf{v}_5 = [0.7, 0.3]^{\mathrm{T}}$. The cluster assignments of the vertices $(z_i)_{i=1}^n$ are drawn from a categorical distribution with probability vector $[1/K, \ldots, 1/K]^{\mathrm{T}}$ and we set $\mathbf{x}_{0i} = \mathbf{v}_{z_i}$, $i \in [n]$. Note that $\mathbf{v}_3$ is very close to the boundary of the parameter space. Let $\mathbf{X}_0 = [\mathbf{x}_{01}, \ldots, \mathbf{x}_{0n}]^{\mathrm{T}}$ and suppose an adjacency matrix $\mathbf{A}$ is generated from $\mathrm{RDPG}(\mathbf{X}_0)$. We consider two sample sizes, $n = 2000$ and $n = 3000$.

We consider the performance of the same estimates as in Section 5.1 given a realization $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X}_0)$: the ASE, the OSE, the MSLE and the BE with the surrogate likelihood computed using the Metropolis–Hastings sampler. For the MSLE, we implement the step-
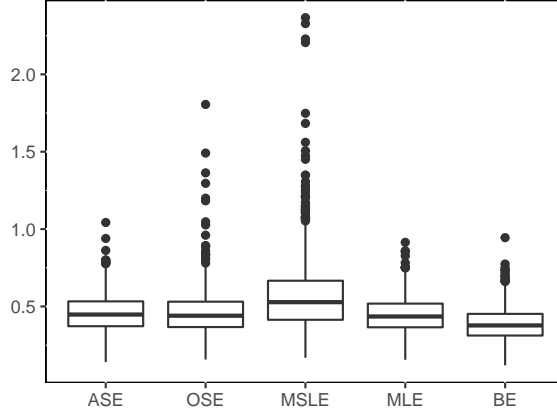
Figure 4: Boxplots of the sum of squared errors (SSEs) $\inf_{\mathbf{W}\in\{\pm 1\}} \|\bar{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$ for Section 5.1 with $n = 30$.

halving stochastic gradient descent algorithm with the batch size set to $s = 500$ and $s = n$ (giving rise to the classical gradient descent algorithm) to compare the computational costs. The setup of the Metropolis–Hastings sampler for the Bayesian estimation is the same as in Section 5.1, and the convergence diagnostics are provided in the Supplementary Material. We take the posterior mean as the point estimate as before. The same experiment is repeated for 1000 independent Monte Carlo replicates.

Similar to Section 5.1, given a generic estimate $\bar{\mathbf{X}}$, we compute the SSEs of the estimates $\inf_{\mathbf{W}\in\mathbb{O}(2)} \|\bar{\mathbf{X}}\mathbf{W} - \mathbf{X}_0\|_{\mathrm{F}}^2$ to measure the estimation accuracy. The summary statistics of these results are visualized in Table 4 (for $n = 2000$) and Table 5 (for $n = 3000$), respectively. We see that the OSE is numerically unstable because v_3 is close to the boundary of the parameter space. Overall, the BE outperforms the other competitors with the least errors, while the ASE and the MSLE have similar performance in terms of the estimation error. The p-values of the pairwise two-sample t-tests among the SSEs of these estimates are reported in Table 6 (for $n = 2000$) and Table 7 (for $n = 3000$), showing that the differences between the BE and the remaining competitors are statistically significant. This phenomenon suggests that, when some latent positions are close to the boundary of the parameter space, the Bayesian estimation method based on the Markov chain Monte Carlo sampler is numerically more stable than the optimization-based frequentist ASE and the MSLE.

The computation times of a single experiment for different estimation procedures are reported in Table 4 and 5. We see that the ASE and the OSE are faster to compute, whereas the MSLE obtained through the classical gradient descent algorithm and the BE are more computationally expensive. We also observe that the stochastic gradient descent algorithm is significantly faster than the classical gradient descent algorithm for finding the MSLE and gains computational efficiency at the cost of estimation accuracy compared to the BE.

The computation times of a single experiment for different estimation procedures are reported in Table 4 and 5. We see that the ASE and the OSE are faster to compute, whereas the MSLE obtained through the classical gradient descent algorithm and the BE are more computationally expensive. We also observe that the stochastic gradient descent algorithm

Estimate	ASE	OSE	MSLE-SGD	MSLE-GD	BE
SSEs	8.570	31.646	8.510	8.513	7.970
Standard errors of SSEs	0.250	18.242	0.250	0.250	0.378
Computation time (seconds)	0.240	0.425	9.920	18.028	98.524

Table 4: Numerical results for Section 5.3: The average sum of squared errors, their standard errors, and the computation time of a single experiment (in seconds). MSLE-SGD and MSLE-GD refer to the MSLE computed by the stochastic gradient descent with batch size being 500 and the classical gradient descent. Sample size $n = 2000$.

Estimate	ASE	OSE	MSLE-SGD	MSLE-GD	BE
SSEs	8.365	44.079	8.323	8.325	7.892
Standard errors of SSEs	0.227	34.972	0.225	0.226	0.410
Computation time (seconds)	0.599	1.127	14.666	28.760	225.676

Table 5: Numerical results for Section 5.3: The average sum of squared errors, their standard errors, and the computation time of a single experiment (in seconds). MSLE-SGD and MSLE-GD refer to the MSLE computed by the stochastic gradient descent with batch size being 500 and the classical gradient descent. Sample size $n = 3000$.

is significantly faster than the classical gradient descent algorithm for finding the MSLE and gains computational efficiency at the cost of estimation accuracy compared to the BE.

5.4 Analysis of the Wikipedia Graph Data Set

In this section, we apply the proposed surrogate likelihood estimation methods to a real-world Wikipedia graph data set. The network data is structured as follows: The vertices represent 1382 Wikipedia articles that are connected to the article named *Algebraic geometry* within two hyperlinks, and an edge is assigned to link two articles if they are connected by a hyperlink. Besides the network itself, each Wikipedia article is also assigned with one of the following six class labels: people, places, dates, things, math and category. The data set is publicly available at at `http://www.cis.jhu.edu/~parky/Data/data.html`.

The goal is to study the clustering accuracy using different estimates when the embedding dimension varies. Given a selected embedding dimension $d \geq 1$, we consider the following four estimates: the ASE, the OSE, the MSLE computed using the step-halving stochastic gradient descent algorithm, and the BE based on the surrogate likelihood (we consider the posterior mean as the point estimate) with the uniform prior distribution on the unit disk for all $\mathbf{x}_i$. Unlike the scenarios in the simulated examples in Sections 5.1 and 5.3, for this real data set, the underlying ground truth of the latent positions is unknown. Rather, only the class labels of the vertices are available to us. To this end, we follow the suggestion in Tang and Priebe (2018) and apply the Gaussian-mixture-model-based clustering to the aforementioned four estimates. Namely, these estimates are regarded as the

Two-sample t-test	BE vs ASE	BE vs OSE	BE vs MSLE-SGD	BE vs MSLE-GD
p-value	6×10^{-28}	9×10^{-17}	3×10^{-24}	2×10^{-24}

Table 6: p-values of the two-sample t-tests between the SSEs of the BE against the remaining estimates for Section 5.3. Sample size $n = 2000$.

Two-sample t-test	BE vs ASE	BE vs OSE	BE vs MSLE-SGD	BE vs MSLE-GD
p-value	1.1×10^{-18}	1.9×10^{-17}	2.3×10^{-16}	1.7×10^{-16}

Table 7: p-values of the two-sample t-tests between the SSEs of the BE against the remaining estimates for Section 5.3. Sample size $n = 3000$.

input for learning the clustering structure of the Wikipedia article network. We report the clustering accuracy using the Rand index (Rand, 1971) as the evaluation metric.

The Rand indices of the clustering results using different estimates across different embedding dimensions $d \in \{1, 2, \ldots, 10\}$ are shown in Table 8. On one hand, we can see that when $d \leq 2$, the adjacency spectral embedding yields better clustering accuracy with a higher Rand index value than the remaining competitors. On the other hand, as the embedding dimension d increases from 2 to 10, the MSLE and the BE with the surrogate likelihood outperform the other two competitors. A plausible explanation of this phenomenon could be that the eigenvectors of the adjacency matrix with smaller eigenvalues are noisier than the top two eigenvectors, but this source of noise is reduced through the additional information introduced by the surrogate likelihood function.

5.5 Analysis of the Political Blogs Network

We now consider the political blogs network (Adamic and Glance, 2005), a benchmark network data that has also been analyzed by Karrer and Newman (2011); Zhao et al. (2012); Amini et al. (2013); Jin (2015); Bickel and Sarkar (2015); Le et al. (2016). The network corresponds to the hyperlinks of blogs regarding U.S. politics after the 2004 presidential election. These blogs are manually classified as either liberal or conservative, which we use as the ground true communities. After following the rule of thumb by extracting the largest connected component and converting the resulting network with undirected edges, we obtain an 1224×1224 adjacency matrix with 33430 edges. We implement the proposed

d	1	2	3	4	5	6	7	8	9	10
ASE	0.745	0.720	0.721	0.723	0.731	0.735	0.736	0.721	0.723	0.715
OSE	0.697	0.706	0.724	0.728	0.735	0.739	0.739	0.741	0.744	0.740
MSLE	0.723	0.711	0.726	0.736	0.739	0.744	0.742	0.744	0.742	0.747
BE	0.718	0.715	0.724	0.735	0.735	0.742	0.743	0.744	0.744	0.745

Table 8: Numerical results of Wikipedia graph data: Rand indices between the class labels and the clustering results based on the four estimates, across embedding dimensions d from 1 to 10, respectively.

Estimate / Clustering method	ASE	OSE	MSLE	BE
Gaussian mixture model	0.1321	0.0416	0.4439	0.4660
Spherical k-means	0.8104	0.8046	0.7726	0.8075
SCORE	0.8193	0.7900	0.8193	0.8134
Computation time (seconds)	0.14	0.16	106.41	54373

Table 9: ARI and the computation times for Section 5.5.

maximum surrogate likelihood estimate and the associated Bayes estimate, together with the adjacency spectral embedding and the one-step estimate as the competitors. We choose the embedding dimension to be $d = 2$ (the same as the number of clusters). Similar to the treatment in Section 5.4, these latent position estimates are then applied to the Gaussian-mixture-model-based clustering, which we compare against the true community labels via the adjusted Rand index (ARI). Since the political blogs network is known to be closer to a degree-corrected stochastic block model (DCSBM) as opposed to the stochastic block model, we also consider the clustering algorithms designed for DCSBM after obtaining the latent position estimates. Specifically, we apply the spherical k-means (Lei and Rinaldo, 2015; Lyzinski et al., 2014) and the spectral clustering on ratios-of-eigenvectors (SCORE) (Jin, 2015). Note that in order to apply SCORE to latent positions estimates, we first compute the left singular vector matrix $\widehat{\mathbf{U}}$ of an estimated latent position matrix $\widehat{\mathbf{X}}$ and then apply SCORE to the orthonormal matrix $\widehat{\mathbf{U}}$. See Table 9 below for the detailed comparison, together with the computation time. When the clustering method is based on the Gaussian mixture model, the proposed Bayes estimate (BE) associated with the surrogate likelihood is more accurate in terms of recovering the liberal-versus-conservative community structure of these political blogs, although its computational cost is much more expensive compared to the maximum surrogate likelihood estimate (MSLE). In this case, both the BE and the MSLE result in significantly better ARI compared to the ASE and the OSE. When the clustering method is either the spherical k-means or SCORE, the ARI differences are marginal. Note that among all different clustering methods applied to different latent position estimates, SCORE applied to the ASE and the MSLE gives the best clustering results compared to other competitors, and they outperform the Gaussian-mixture-model-based clustering by a large margin.

6. Discussion

In this paper, we propose a novel surrogate likelihood estimation framework for random dot product graphs. The surrogate likelihood has several fascinating properties, including the separability and the log-concavity, that facilitate theoretical analysis and practical computation. We study both the maximum surrogate likelihood estimation from the frequentist perspective and the Bayesian estimation using the surrogate likelihood. In particular, we establish the existence, uniqueness, and asymptotic normality of the maximum likelihood estimator, and propose a convenient stochastic gradient descent algorithm for the computation. Furthermore, we derive the $O(n^{-1})$ biases of the maximum surrogate likelihood

estimator and the one-step estimator. These formulae illustrate how the former outperforms the latter in finite sample problems. We also establish the Bernstein–von Mises theorem of the posterior distribution with the surrogate likelihood function and show that the resulting credible sets have the correct frequentist coverage probabilities. It turns out that the maximum surrogate likelihood estimator and the Bayes estimator are asymptotically efficient in the sense of local efficiency (Xie and Xu, 2023), and they outperform the baseline adjacency spectral embedding in terms of smaller asymptotic mean squared errors. Our numerical examples also suggest that the proposed surrogate likelihood methodology is more favorable than the previously developed one-step estimator (Xie and Xu, 2023) in some finite sample problems. In particular, we have observed that for networks with comparatively small and moderate sizes, the empirical improvement of the maximum surrogate likelihood estimates over the one-step estimates, among other competitors, is more significant. Intuitively, such a phenomenon can be partially explained by the $O(n^{-1})$ bias derived in Theorem 7 and Example 3. When n is small or moderate, the $O(n^{-1})$ bias difference is more observable as opposed to the case when n is large. For example, in Section 5, the outperformance of the maximum surrogate likelihood estimates over the one-step estimates is more significant when $n = 30$ (Table 2) and 300 (Table 3) than when $n \in \{1000, 2000, 3000\}$.

Our current methodology and theory are designed for random dot product graphs with positive semidefinite edge probability matrices. These networks can only model the so-called assortative mixing networks and exclude many interesting examples, such as disassortative mixing stochastic block models with larger between-community connection probabilities and smaller within-community connection probabilities. Extending the current framework to random graphs with possibly indefinite edge probability matrices is straightforward by introducing the generalized random dot product graphs (Rubin-Delanchy et al., 2022), where $\mathbb{E}A_{ij} = \mathbf{x}_i^{\mathrm{T}}\mathbf{I}_{p,q}\mathbf{x}_j$ and $\mathbf{I}_{p,q}$ is a diagonal matrix whose first p diagonals are $+1$ and the remaining q diagonals -1 with $d = p + q$. The generalized random dot product graphs allow disassortative mixing networks. The trick is to replace the ASE $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}}$ with the sign-adjusted ASE $\widetilde{\mathbf{Y}} = [\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_n]^{\mathrm{T}} = [|\widehat{\lambda}_1|\widehat{\mathbf{u}}_1, \ldots, |\widehat{\lambda}_d|\widehat{\mathbf{u}}_d]$, where $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d$ are the largest d-eigenvalues of $\mathbf{A}$ in absolute value and $\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_d$ are the associated eigenvectors. The associated theory and computation methods can be easily extended accordingly.

In Section 2.3, we have seen that the one-step estimator also corresponds to the maximizer of an approximate likelihood function, but it has a worse approximation quality than the proposed surrogate likelihood near the oracle maximum likelihood estimator. Surprisingly, under a framework of generalized estimating equations proposed by Xie and Wu (2024), the gradients of both the surrogate log-likelihood function and the approximate log-likelihood function associated with the one-step estimator can be viewed as some generalized estimating equations that take advantage of the likelihood function information. This intuition conforms to the fact that the estimators based on the approximation of likelihood are asymptotically equivalent up to the first order. We have also found in some finite sample problems that the maximum surrogate likelihood estimator outperforms the one-step estimator, which can be explained by the difference of their $O(n^{-1})$ bias. However, a more systematic way to study the performance difference of these two estimators requires the analysis of $O(n^{-3/2})$ mean squared errors by following the spirit of Pfanzagl and Wefelmeyer (1978) and Newey and Smith (2004), which in turn requires the analysis

of $O(n^{-3/2})$ behavior of the ASE beyond Xie and Zhang (2025). This is an interesting direction that we defer to future research.

Appendix A. Preliminary Results for the proofs

Lemma 13 *Let* $\mathbf{A} \sim \mathrm{RDPG}(\rho_n^{1/2}\mathbf{X}_0)$ *with* $n\rho_n \gtrsim \log n$. *Denote by* $\boldsymbol{\Delta}_n = (1/n)\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$. *Assume* $\lambda_d(\boldsymbol{\Delta}_n) \geq \lambda$ *for some constant* $\lambda > 0$ *for all sufficiently large* n, *and* $\min_{i,j \in [n]}(\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}, 1 - \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}) \geq \delta$ *for some constant* $\delta > 0$. *Then for all* $c > 0$, *there exists a constant* $N_{c,\lambda} \in \mathbb{N}_+$ *depending on* c, λ, *such that for all* $n \geq N_{c,\lambda}$,

$$\|\widetilde{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{2\to\infty} \lesssim_{c,\lambda} \sqrt{\frac{\log n}{n}}.$$

with probability at least $1 - n^{-c}$.

Proof Denote by $\kappa(\boldsymbol{\Delta}_n) = \lambda_1(\boldsymbol{\Delta}_n)/\lambda_d(\boldsymbol{\Delta}_n)$. By Corollary 4.1 in Xie (2024), for all $c > 0$, we can pick a constant $N_c \in \mathbb{N}_+$ such that for all $n \geq N_c$, with probability at least $1 - n^{-c}$,

$$\|\widetilde{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{2\to\infty} \lesssim_c \frac{\|\mathbf{U}_{\mathbf{P}}\|_{2\to\infty}}{(n\rho_n)^{1/2}\lambda_d(\boldsymbol{\Delta}_n)^2} \max\left\{\frac{(\log n)^{1/2}}{\lambda_d(\boldsymbol{\Delta}_n)^2}, \frac{\kappa(\boldsymbol{\Delta}_n)}{\lambda_d(\boldsymbol{\Delta}_n)^2}, \log n\right\}$$
$$+ \frac{(\log n)^{1/2}\|\mathbf{U}_{\mathbf{P}}\|_{2\to\infty}}{\lambda_d(\boldsymbol{\Delta}_n)^{1/2}}.$$

Observe that $\lambda_d(\boldsymbol{\Delta}_n)$ is lower bounded by a constant $\lambda > 0$ for sufficiently large n, and $\lambda_1(\boldsymbol{\Delta}_n) \leq (1/n)\|\mathbf{X}_0\|_{\mathrm{F}}^2 \leq 1$. Also note that

$$\|\mathbf{U}_{\mathbf{P}}\|_{2\to\infty} \leq \|\rho_n^{1/2}\mathbf{X}_0\|_{2\to\infty}\|\mathbf{S}_{\mathbf{P}}^{-1/2}\|_2 \leq \sqrt{\frac{\rho_n}{n\rho_n\lambda_d(\boldsymbol{\Delta}_n)}} \leq \frac{1}{\sqrt{n\lambda}}.$$

Therefore, by the fact that $(\log n)/(n\rho_n)$ is bounded, we can pick a constant $N_{c,\lambda} \in \mathbb{N}_+$ depending on c, λ, such that for all $n \geq N_{c,\lambda}$, with probability at least $1 - n^{-c}$,

$$\|\widetilde{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{2\to\infty} \lesssim_c \|\mathbf{U}_{\mathbf{P}}\|_{2\to\infty}\frac{\log n}{(n\rho_n)^{1/2}\lambda^5} + \|\mathbf{U}_{\mathbf{P}}\|_{2\to\infty}\frac{(\log n)^{1/2}}{\lambda^{1/2}} \lesssim_{c,\lambda} \sqrt{\frac{\log n}{n}}.$$

This completes the proof. ∎

Lemma 14 (Some frequently used results) *Suppose* $\mathbf{A} \sim \mathrm{RDPG}(\rho_n^{1/2}\mathbf{X}_0)$ *and assume the conditions of Theorem 3 hold. Denote by* $\widetilde{p}_{ij} = \widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j$, $i, j \in [n]$. *Then for all* $c > 0$, *there exists a constant* $N_{c,\delta,\lambda} \in \mathbb{N}_+$ *depending on* c, δ, λ *such that for all* $n \geq N_{c,\delta,\lambda}$, *the following hold with probability at least* $1 - n^{-c}$:

$$\max_{j \in [n]}\|\widetilde{\mathbf{x}}_j\|_2 \leq^{1/2} (1 - \frac{\delta}{2})\rho_n,$$

$$\max_{i,j \in [n]}|\widetilde{p}_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}| \lesssim_{c,\lambda} \rho_n^{1/2}\sqrt{\frac{\log n}{n}},$$

$$\frac{\delta}{2}\rho_n \leq \min_{i,j\in[n]} \widetilde{p}_{ij} \leq \max_{i,j\in[n]} \widetilde{p}_{ij} \leq (1 - \frac{\delta}{2})\rho_n,$$

$$\max_{j\in[n]} \|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W} - \rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\|_2 \lesssim_{c,\lambda} \rho_n^{1/2}\sqrt{\frac{\log n}{n}}.$$

Proof For the first result, by Lemma 13 and the condition that $\frac{\log n}{n\rho_n} \to 0$, we can pick a constant $N_{c,\delta,\lambda} \in \mathbb{N}_+$ depending on c,δ,λ such that for all $n \geq N_{c,\delta,\lambda}$, with probability at leat $1 - n^{-c}$,

$$\|\widetilde{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{2\to\infty} = \max_{j\in[n]} \|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \rho_n^{1/2}\mathbf{x}_{0j}\|_2 \leq \rho_n^{1/2}\left(1 - \frac{\delta}{2} - \sqrt{1-\delta}\right).$$

This is because $(1 - \delta/2)^2 = 1 - \delta + \delta^2/4 > 1 - \delta$. Then

$$\max_{j\in[n]} \|\widetilde{\mathbf{x}}_j\|_2 \leq \max_{j\in[n]} \|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \rho_n^{1/2}\mathbf{x}_{0j}\|_2 + \max_{j\in[n]} \|\rho_n^{1/2}\mathbf{x}_{0j}\|_2$$

$$\leq \rho_n^{1/2}\left(1 - \frac{\delta}{2} - \sqrt{1-\delta}\right) + \rho_n^{1/2}\sqrt{1-\delta}.$$

For the second result, over the same event as above, we have

$$\max_{i,j\in[n]} |\widetilde{p}_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}| \leq \max_{i,j\in[n]} |\widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \rho_n^{1/2}\mathbf{x}_{0j})| + \max_{i,j\in[n]} |(\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i})^{\mathrm{T}}\rho_n^{1/2}\mathbf{x}_{0j}|$$

$$\leq (\max_{j\in[n]} \|\widetilde{\mathbf{x}}_j\|_2 + \rho_n^{1/2})\|\widetilde{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{2\to\infty} \lesssim_{c,\lambda} \rho_n^{1/2}\sqrt{\frac{\log n}{n}}.$$

For the third result, over the same event as above, we have

$$\max_{i,j\in[n]} \widetilde{p}_{ij} \leq \max_{i,j\in[n]} |\widetilde{p}_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}| + \max_{i,j\in[n]} \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j} \leq C_{c,\lambda}\rho_n^{1/2}\sqrt{\frac{\log n}{n}} + \rho_n(1-\delta).$$

Since $\frac{\log n}{n\rho_n} \to 0$ and $\max_{i,j\in[n]} \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j} \leq 1 - \delta$, we can pick a (possibly larger) constant $N_{c,\delta,\lambda}$ such that $C_{c,\lambda}^{1/2}\sqrt{\frac{\log n}{n\rho_n}} \leq \delta/2$ for all $n \geq N_{c,\delta,\lambda}$. Then

$$\max_{i,j\in[n]} \widetilde{p}_{ij} \leq \rho_n(1 - \frac{\delta}{2}).$$

Similarly,

$$\min_{i,j\in[n]} \widetilde{p}_{ij} \geq \min_{i,j\in[n]} |\widetilde{p}_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}| - \max_{i,j\in[n]} \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j} \geq \rho_n\delta - C_{c,\lambda}\rho_n^{1/2}\sqrt{\frac{\log n}{n}} \geq \frac{\rho_n\delta}{2}.$$

For the fourth one, over the same event as above, we have

$$\max_{j\in[n]} \|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W} - \rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\|_2$$

$$\leq \max_{j\in[n]} \|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j(\widetilde{\mathbf{x}}_j\mathbf{W} - \rho_n^{1/2}\mathbf{x}_{0j}^{\mathrm{T}})\|_2 + \max_{j\in[n]} \|(\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \rho_n^{1/2}\mathbf{x}_{0j})\rho_n^{1/2}\mathbf{x}_{0j}^{\mathrm{T}}\|_2$$

$$\leq (\max_{j\in[n]} \|\widetilde{\mathbf{x}}_j\|_2 + \rho_n^{1/2})\|\widetilde{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{2\to\infty} \lesssim_{c,\lambda} \rho_n^{1/2}\sqrt{\frac{\log n}{n}}.$$

∎

Lemma 15 (Concentration of Hessian matrices) *Suppose* $\mathbf{A} \sim \mathrm{RDPG}(\rho_n^{1/2}\mathbf{X}_0)$ *and assume the conditions of Theorem 3 hold. Denote by* $\widetilde{p}_{ij} = \widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j$, $i, j \in [n]$ *and let* $\epsilon > 0$ *be sufficiently small. Then for all* $c > 0$, *there exists a constant* $N_{c,\delta,\lambda} \in \mathbb{N}_+$ *depending on* c, δ, λ *such that for all* $n \geq N_{c,\delta,\lambda}$, *the following hold with probability at least* $1 - n^{-c}$:

$$\max_{i\in[n]} \sup_{\mathbf{x}_i:\|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i-\rho_n^{0.5}\mathbf{x}_{0i}\|_2\leq\epsilon} \left\|\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{\widetilde{p}_{ij}} + \frac{1-A_{ij}}{(1-\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)^2}\right\}\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W}\right.$$
$$\left.- \frac{1}{n}\sum_{j=1}^{n}\frac{\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}\right\|_2 \quad \lesssim_{c,\delta,\lambda} \rho_n^{1.5}\epsilon_n + \sqrt{\frac{\log n}{n\rho_n}},$$

$$\left\|\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{\widetilde{p}_{ij}} + \frac{1-A_{ij}}{(1-\widetilde{p}_{ij})^2}\right\}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\widetilde{p}_{ij}(1-\widetilde{p}_{ij})}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\right\|_2 \lesssim_{c,\delta,\lambda} \rho_n^{1.5}\sqrt{\frac{\log n}{n}}.$$

Proof For simplicity of notation, denote by $p_{0ij} = \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}$. The large probability bounds below are with regard to $n \geq N_{c,\delta,\lambda}$ for some large constant $N_{c,\delta,\lambda}$ depending on c, δ, λ. ∎ We show the first conclusion first. Write

$$\left\|\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{\widetilde{p}_{ij}} + \frac{1-A_{ij}}{(1-\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)^2}\right\}\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W} - \frac{1}{n}\sum_{j=1}^{n}\frac{\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}\right\|_2$$

$$\leq \left\|\frac{1}{n}\sum_{j=1}^{n}(1-A_{ij})\left\{\frac{1}{(1-\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)^2} - \frac{1}{(1-p_{0ij})^2}\right\}\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W}\right\|_2$$

$$+ \left\|\frac{1}{n}\sum_{j=1}^{n}\frac{A_{ij}-p_{0ij}}{(1-p_{0ij})^2}\left(\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W} - \rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\right)\right\|_2 + \left\|\frac{1}{n}\sum_{j=1}^{n}\frac{A_{ij}-p_{0ij}}{(1-p_{0ij})^2}\rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\right\|_2$$

$$+ \left\|\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{\widetilde{p}_{ij}} - \frac{1}{p_{0ij}}\right\}\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W}\right\|_2 + \left\|\frac{1}{n}\sum_{j=1}^{n}\frac{\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W} - \rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}}{p_{0ij}(1-p_{0ij})}\right\|_2.$$

For the first term, with probability at least $1 - n^{-c}$,

$$\max_{i\in[n]} \sup_{\mathbf{x}_i:\|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i-\rho_n^{1/2}\mathbf{x}_{0i}\|_2\leq\epsilon} \left\|\frac{1}{n}\sum_{j=1}^{n}(1-A_{ij})\left\{\frac{1}{(1-\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)^2} - \frac{1}{(1-p_{0ij})^2}\right\}\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W}\right\|_2$$

$$\leq \max_{i\in[n]} \sup_{\mathbf{x}_i:\|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i-\rho_n^{1/2}\mathbf{x}_{0i}\|_2\leq\epsilon} \frac{1}{n}\sum_{j=1}^{n}2\frac{|(\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j - p_{0ij})(2 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j - p_{0ij})|}{(1-\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)^2(1-p_{0ij})^2}\|\widetilde{\mathbf{x}}_j\|_2^2$$

$$\lesssim_{c,\delta,\lambda} \max_{i\in[n]} \sup_{\mathbf{x}_i:\|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i-\rho_n^{1/2}\mathbf{x}_{0i}\|_2\le\epsilon} \rho_n \frac{1}{n} \sum_{j=1}^{n} \left|\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j - p_{0ij}\right|$$

$$\le \max_{i\in[n]} \sup_{\mathbf{x}_i:\|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i-\rho_n^{1/2}\mathbf{x}_{0i}\|_2\le\epsilon} \rho_n \frac{1}{n} \sum_{j=1}^{n} \left\{ \left\|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i - \rho_n^{1/2}\mathbf{x}_{0i}\right\|_2 \|\widetilde{\mathbf{x}}_j\|_2 \right.$$
$$\left. + \|\rho_n^{1/2}\mathbf{x}_{0i}\|_2 \left\|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \rho_n^{1/2}\mathbf{x}_{0j}\right\|_2 \right\}$$

$$\lesssim_{c,\delta,\lambda} \rho_n^{\frac{3}{2}}\epsilon + \rho_n^{\frac{3}{2}}\sqrt{\frac{\log n}{n}},$$

where in the second inequality we use Lemma 14, in the third inequality triangle inequality and Cauchy–Schwarz inequality, and in the fourth inequality Lemma 13 and Lemma 14. For the second term, with probability at least $1 - n^{-c}$,

$$\max_{i\in[n]} \left\|\frac{1}{n} \sum_{j=1}^{n} \frac{A_{ij} - p_{0ij}}{(1 - p_{0ij})^2} \left(\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W} - \rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\right)\right\|_2$$

$$\lesssim_{\delta} \frac{1}{n} \left\|\mathbf{A} - \rho_n\mathbf{X}_0\mathbf{X}_0^{\mathrm{T}}\right\|_\infty \max_{j\in[n]} \left\|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W} - \rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\right\|_2$$

$$\le \frac{1}{n}\left(\|\mathbf{A}\|_\infty + \|\rho_n\mathbf{X}_0\mathbf{X}_0^{\mathrm{T}}\|_\infty\right) \max_{j\in[n]} \left\|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W} - \rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\right\|_2 \lesssim_{c,\delta,\lambda} \rho_n^{\frac{3}{2}}\sqrt{\frac{\log n}{n}},$$

by Lemma 14 and the result that $\|\mathbf{A}\|_\infty \lesssim_c n\rho_n$ with probability at least $1 - n^{-c}$ which follows from triangle inequality and Bernstein's inequality.

For the third term, for a typical (k,l)th entry, by Bernstein's inequality and a union bound over $i \in [n]$, for all $t > 0$,

$$\mathbb{P}\left\{\max_{i\in[n]} \left|\frac{1}{n}\sum_{j=1}^{n}(A_{ij} - p_{0ij})\frac{\rho_n x_{0jk}x_{0jl}}{(1-p_{0ij})^2}\right| \ge t\right\}$$

$$\le 2n\exp\left\{\frac{-3n^2t^2}{6\sum_{j=1}^{n}\frac{\rho_n^2 x_{0jk}^2 x_{0jl}^2}{(1-p_{0ij})^4}p_{0ij}(1-p_{0ij}) + 2\max_{j\in[n]}\frac{\rho_n x_{0jk}x_{0jl}}{(1-p_{0ij})^2}nt}\right\}$$

$$\le 2n\exp\left\{-K_\delta\frac{nt^2}{\rho_n^3 + \rho_n t}\right\},$$

where $K_\delta > 0$ is a constant depending on δ. Taking $t = C\sqrt{(\rho_n^3\log n)/n}$ for an appropriate constant $C > 0$, we see that

$$\max_{i\in[n]}\left|\frac{1}{n}\sum_{j=1}^{n}(A_{ij}-p_{0ij})\frac{\rho_n x_{0jk}x_{0jl}}{(1-p_{0ij})^2}\right| \lesssim_{c,\delta} \sqrt{\frac{\rho_n^3\log n}{n}}$$

with probability at least $1 - n^{-c}$. Since d is fixed (it implicitly depends on λ), we have

$$\max_{i\in[n]}\left\|\frac{1}{n}\sum_{j=1}^{n}(A_{ij}-p_{0ij})\frac{\rho_n\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}}{(1-p_{0ij})^2}\right\|_2 \lesssim_{c,\delta,\lambda} \rho_n^{\frac{3}{2}}\sqrt{\frac{\log n}{n}}$$

27

with probability at least $1 - n^{-c}$.

For the fourth term, with probability at least $1 - n^{-c}$,

$$
\max_{i \in [n]} \left\| \frac{1}{n} \sum_{j=1}^{n} \left\{ \frac{1}{\widetilde{p}_{ij}} - \frac{1}{p_{0ij}} \right\} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W} \right\|_2 \leq \max_{i,j \in [n]} \frac{|\widetilde{p}_{ij} - p_{0ij}|}{\widetilde{p}_{ij} p_{0ij}} \|\widetilde{\mathbf{x}}_j\|_2^2 \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}
$$

by Lemma 14.

For the fifth term, with probability at least $1 - n^{-c}$,

$$
\max_{i \in [n]} \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W} - \rho_n \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}}}{p_{0ij}(1 - p_{0ij})} \right\|_2 \lesssim_\delta \rho_n^{-1} \max_{j \in [n]} \|\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W} - \rho_n \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}}\|_2
$$

$$
\lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}
$$

by Lemma 14. So the first conclusion is shown by combining the above five bounds.
■ Next, we show the second conclusion. Write

$$
\left\| \frac{1}{n} \sum_{j=1}^{n} \left\{ \frac{1}{\widetilde{p}_{ij}} + \frac{(1 - A_{ij})}{(1 - \widetilde{p}_{ij})^2} \right\} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} - \frac{1}{n} \sum_{j=1}^{n} \frac{\widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}}}{\widetilde{p}_{ij}(1 - \widetilde{p}_{ij})} \right\|_2
$$

$$
= \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{(A_{ij} - \widetilde{p}_{ij}) \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W}}{(1 - \widetilde{p}_{ij})^2} \right\|_2
$$

$$
\leq \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}) \rho_n \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}}}{(1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j})^2} \right\|_2
$$

$$
+ \left\| \frac{1}{n} \sum_{j=1}^{n} (A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}) \left\{ \frac{\rho_n \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}}}{(1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j})^2} - \frac{\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W}}{(1 - \widetilde{p}_{ij})^2} \right\} \right\|_2
$$

$$
+ \left\| \frac{1}{n} \sum_{j=1}^{n} (\widetilde{p}_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}) \frac{\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W}}{(1 - \widetilde{p}_{ij})^2} \right\|_2.
$$

The first term is $O(\rho_n^{\frac{3}{2}} \sqrt{\frac{\log n}{n}})$ with probability at least $1 - n^{-c}$ as previously shown.
For the second term, with probability at least $1 - n^{-c}$,

$$
\left\| \frac{1}{n} \sum_{j=1}^{n} (A_{ij} - p_{0ij}) \left\{ \frac{\rho_n \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}}}{(1 - p_{0ij})^2} - \frac{\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W}}{(1 - \widetilde{p}_{ij})^2} \right\} \right\|_2
$$

$$
\leq \frac{1}{n} \sum_{j=1}^{n} |A_{ij} - p_{0ij}| \left\{ \left\| \frac{\rho_n \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}}}{(1 - p_{0ij})^2} - \frac{\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W}}{(1 - p_{0ij})^2} \right\|_2 + \left\| \frac{\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W}}{(1 - p_{0ij})^2} - \frac{\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}} \mathbf{W}}{(1 - \widetilde{p}_{ij})^2} \right\|_2 \right\}
$$

$$
\leq \frac{1}{n} \|\mathbf{A} - \rho_n \mathbf{X}_0 \mathbf{X}_0^{\mathrm{T}}\|_\infty
$$

28

$$\times \max_{i,j\in[n]} \left\{ \frac{\|\rho_n \mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}} - \mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W}\|_2}{(1-p_{0ij})^2} + \left| \frac{(\widetilde{p}_{ij}-p_{0ij})(2-\widetilde{p}_{ij}-p_{0ij})}{(1-p_{0ij})^2(1-\widetilde{p}_{ij})^2} \right| \|\widetilde{\mathbf{x}}_j\|_2^2 \right\}$$

$$\lesssim_{c,\delta,\lambda} \frac{1}{n}\rho_n \left(\rho_n^{1/2}\sqrt{\frac{\log n}{n}} + \rho_n^{1/2}\sqrt{\frac{\log n}{n}}\rho_n \right)$$

$$\lesssim_{c,\delta,\lambda} \rho_n^{\frac{3}{2}}\sqrt{\frac{\log n}{n}}$$

by Cauchy–Schwarz inequality, Lemma 14, and $\|\mathbf{A}-\rho_n\mathbf{X}_0\mathbf{X}_0^{\mathrm{T}}\|_\infty \leq \|\mathbf{A}\|_\infty + \|\rho_n\mathbf{X}_0\mathbf{X}_0^{\mathrm{T}}\|_\infty \lesssim_c n\rho_n$ with probability at least $1-n^{-c}$.

For the third term, with probability at least $1-n^{-c}$,

$$\left\| \frac{1}{n}\sum_{j=1}^n (\widetilde{p}_{ij}-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}) \frac{\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{W}}{(1-\widetilde{p}_{ij})^2} \right\|_2 \leq \max_{i,j\in[n]} |\widetilde{p}_{ij}-p_{0ij}| \cdot \max_{i,j\in[n]} \frac{1}{(1-\widetilde{p}_{ij})^2} \cdot \max_{j\in[n]} \|\widetilde{\mathbf{x}}_j\|_2^2$$

$$\lesssim_{c,\delta,\lambda} \rho_n^{\frac{3}{2}}\sqrt{\frac{\log n}{n}}$$

by Lemma 14. So the second conclusion is shown by combining the above three bounds. ∎

Lemma 16 *Suppose* $\mathbf{A} \sim \mathrm{RDPG}(\rho_n^{1/2}\mathbf{X}_0)$ *and assume the conditions of Theorem 4 hold. Denote by*

$$Z = Z(\mathbf{A}) = \sum_{i=1}^n \left\| \frac{1}{n\rho_n^{1/2}}\sum_{j=1}^n \frac{(A_{ij}-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} \right\|_2^2.$$

Then $Z = \mathbb{E}_0 Z + o_{\mathbb{P}_0}(1)$.

Proof Denote by

$$\gamma_{ij} = \frac{\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j}}{n\rho_n^{1/2}\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1-\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}, \quad i,j\in[n].$$

Then we have

$$Z - \mathbb{E}_0 Z = \sum_{i=1}^n \sum_{a=1}^n \sum_{b=1}^n E_{ia}E_{ib}\gamma_{ia}^{\mathrm{T}}\gamma_{ib}\mathbb{1}(a\neq b)$$

$$= \sum_{i=1}^n \sum_{a\geq i} \sum_{b\geq i} E_{ia}E_{ib}\gamma_{ia}^{\mathrm{T}}\gamma_{ib}\mathbb{1}(a\neq b) + \sum_{i=1}^n \sum_{a<i} \sum_{b\geq i} E_{ai}E_{ib}\gamma_{ia}^{\mathrm{T}}\gamma_{ib}\mathbb{1}(a\neq b)$$

$$+ \sum_{i=1}^n \sum_{a\geq i} \sum_{b<i} E_{ia}E_{bi}\gamma_{ia}^{\mathrm{T}}\gamma_{ib}\mathbb{1}(a\neq b) + \sum_{i=1}^n \sum_{a<i} \sum_{b<i} E_{ai}E_{bi}\gamma_{ia}^{\mathrm{T}}\gamma_{ib}\mathbb{1}(a\neq b).$$

To prove the desired result, we need the following decoupling inequality for U-statistic-type random variables.

Lemma 17 (Theorem 1 in de la Pena and Montgomery-Smith (1995)) *Let $(X_i)_i$ be a sequence of independent random variables on a measurable space and let $(X_i^{(1)})_i, (X_i^{(2)})_i$ be two independent copies of $(X_i)_i$. Let $f_{i_1 i_2}$ be families of functions of k variables taking $(S \times S)$ into a Banach space $(B, \|\cdot\|_2)$. Then, for all $n \geq 2$, $t > 0$, there exists a numerical constant C such that*

$$\mathbb{P}\left\{ \left\| \sum_{1 \leq i_1 \neq i_2 \leq n} f_{i_1 i_2}(X_{i_1}^{(1)}, X_{i_2}^{(1)}) \right\| \geq t \right\} \leq C\mathbb{P}\left\{ C \left\| \sum_{1 \leq i_1 \neq i_2 \leq n} f_{i_1 i_2}(X_{i_1}^{(1)}, X_{i_2}^{(2)}) \right\|_2 \geq t \right\}$$

Now we apply Lemma 17 with $(X_i)_i = (E_{ij} : 1 \leq i \leq j \leq n)$, and $f_{(i_1,a),(i_2,b)}(E_{i_1 a}, E_{i_2 b}) = \mathbb{1}(i_1 = i_2)\gamma_{i_1 a}^{\mathrm{T}}\gamma_{i_2 b}$, $f_{(i_1,a),(i_2,b)}(E_{i_1 a}, E_{i_2 b}) = \mathbb{1}(a = i_2)\gamma_{a i_1}^{\mathrm{T}}\gamma_{i_2 b}$, $f_{(i_1,a),(i_2,b)}(E_{i_1 a}, E_{i_2 b}) = \mathbb{1}(i_1 = b)\gamma_{i_1 a}^{\mathrm{T}}\gamma_{b i_2}$, and $f_{(i_1,a),(i_2,b)}(E_{i_1 a}, E_{i_2 b}) = \mathbb{1}(a = b)\gamma_{a i_1}^{\mathrm{T}}\gamma_{b i_2}$, for the four terms, respectively. Without loss of generality, it is sufficient to work with the first term. Now let $(\bar{E}_{ij} : 1 \leq i \leq j \leq n)$ be an independent copy of $(E_{ij} : 1 \leq i \leq j \leq n)$. It is sufficient to show that

$$\sum_{i=1}^{n} \sum_{a \geq i} \sum_{b \geq i} E_{ia} \bar{E}_{ib} \gamma_{ia}^{\mathrm{T}} \gamma_{ib} \mathbb{1}(a \neq b) = o_{\mathbb{P}_0}(1).$$

By Bernstein's inequality and the independence between E_{ia} and $\bar{E}_{ib}$, for all $c > 0$, there exists a constant $K_c > 0$, such that

$$\left| \sum_{i=1}^{n} \sum_{a \geq i} \sum_{b \geq i} E_{ia} \bar{E}_{ib} \gamma_{ia}^{\mathrm{T}} \gamma_{ib} \mathbb{1}(a \neq b) \right| \leq K_c n (\rho_n \log n)^{1/2} \max_{1 \leq i, a \leq n} \left| \sum_{b=1}^{n} \bar{E}_{ib} \gamma_{ia}^{\mathrm{T}} \gamma_{ib} \right| \leq \frac{K_c^2 \log n}{\sqrt{n}}$$

with probability at least $1 - O(n^{-c})$. The proof is thus completed. ∎

Theorem 18 (Theorem 4.7 in Xie, 2024) *Suppose $\mathbf{A} \sim \mathrm{RDPG}(\rho_n^{1/2}\mathbf{X}_0)$ and assume the conditions of Theorem 3 hold. Define the one-step estimator $\widehat{\mathbf{x}}_i^{(\mathrm{OS})}$ by*

$$\widehat{\mathbf{x}}_i^{(\mathrm{OS})} = \widetilde{\mathbf{x}}_i + \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{\widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}}}{\widetilde{p}_{ij}(1 - \widetilde{p}_{ij})} \right\}^{-1} \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{(A_{ij} - \widetilde{p}_{ij})\widetilde{\mathbf{x}}_j}{\widetilde{p}_{ij}(1 - \widetilde{p}_{ij})} \right\}.$$

Then

$$\mathbf{G}_{0in}^{1/2}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \rho_n^{1/2}\mathbf{x}_{0i}) = \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{G}_{0in}^{-1/2}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} + \mathbf{r}_{in}^{(\mathrm{OS})},$$

where

$$\mathbf{G}_{0in} = \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})},$$

and for all $c > 0$, there exists a constant $N_{c,\delta,\lambda} \in \mathbb{N}_+$ depending on c, δ, λ, such that for all $n \geq N_{c,\delta,\lambda}$, with probability at least $1 - (n\rho_n)^{-c}$, $\|\mathbf{r}_{in}^{(\mathrm{OS})}\|_2 \lesssim (\log(n\rho_n))^2/(n\rho_n^{1/2})$. Furthermore,

$$\sqrt{n}\mathbf{G}_{0in}^{1/2}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d),$$

Appendix B. Proofs of the Main Results

B.1 Proof of Theorem 3

Proof ■ We first prove existence. For all $c > 0$, there exists $N_{c,\delta,\lambda} \in \mathbb{N}_+$ such that

$$\sup_{\|\mathbf{x}_i\|_2 \leq 1} \max_{j \in [n]} |\mathbf{x}_i^T \widetilde{\mathbf{x}}_j| \leq \max_{j \in [n]} \|\widetilde{\mathbf{x}}_j\|_2 \leq \rho_n(1 - \frac{\delta}{2}) < 1$$

with probability at least $1 - n^{-c}$, where the first inequality follows from Cauchy–Schwarz inequality, the second from Lemma 14. By definition of $\widetilde{M}_{in}(\mathbf{x}_i)$, it is continuous over the closed unit ball $\{\mathbf{x}_i \in \mathbb{R}^d : \|\mathbf{x}_i\|_2 \leq 1\}$ over this event. Hence the maximizer $\widehat{\mathbf{x}}_i$ of $\widetilde{M}_{in}(\mathbf{x}_i)$ exists with probability at least $1 - n^{-c}$.

■ Next we prove uniqueness. By definition, with probability at least $1 - n^{-c}$, $\widetilde{M}_{in}(\mathbf{x}_i)$ is twice continuously differentiable, with

$$-\frac{\partial \widetilde{M}_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^T}(\mathbf{x}_i) = \frac{1}{n} \sum_{j=1}^n \left\{ \frac{1}{\widetilde{p}_{ij}} + \frac{1 - A_{ij}}{(1 - \mathbf{x}_i^T \widetilde{\mathbf{x}}_j)^2} \right\} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^T \succeq \frac{1}{n} \sum_{j=1}^n \frac{\widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^T}{\widetilde{p}_{ij}}$$

$$\succeq \frac{1}{n\rho_n} \sum_{j=1}^n \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} \succeq \frac{1}{n\rho_n} \sigma_d(\widetilde{\mathbf{X}})^2 \mathbf{I}_d.$$

By Theorem 5.2 in Lei and Rinaldo (2015) and Weyl's inequality, there exists a constant depending on c, such that with probability at least $1 - n^{-c}$,

$$\sigma_d(\widetilde{\mathbf{X}})^2 = \lambda_d(\mathbf{A}) \geq \frac{1}{2}\lambda_d(\rho_n \mathbf{X}_0 \mathbf{X}_0^T) = \frac{1}{2}n\rho_n \lambda_d\left(\frac{1}{n}\mathbf{X}_0 \mathbf{X}_0^T\right) \geq \frac{1}{2}n\rho_n \lambda > 0.$$

Therefore, for all $c > 0$, there exists a constant $N_{c,\delta,\lambda} \in \mathbb{N}_+$ such that for all $n \geq N_{c,\delta,\lambda}$, with probability at least $1 - n^{-c}$, $\widetilde{M}_{in}(\mathbf{x}_i)$ is strictly concave. Hence it has a unique maximizer $\widehat{\mathbf{x}}_i$. ■

B.2 Proof of Theorem 4

Proof ■ We first establish the following consistency result: For all $c > 0$, there exists a constant $N_{c,\delta,\lambda} \in \mathbb{N}_+$ depending on c, δ, λ such that for all $n \geq N_{c,\delta,\lambda} \in \mathbb{N}_+$, there exists an orthogonal matrix $\mathbf{W} \in \mathbb{O}(d)$, such that with probability at least $1 - n^{-c}$,

$$\max_{i \in [n]} \|\mathbf{W}^T \widehat{\mathbf{x}}_i - \rho_n^{1/2} \mathbf{x}_{0i}\|_2 \lesssim_{c,\delta,\lambda} \{\log n / (n\rho_n)\}^{1/2}.$$

Define $\widetilde{M}_{in}(\mathbf{x}_i) = (1/n)\widetilde{\ell}_{in}(\mathbf{x}_i)$ and the population counterpart of $\widetilde{M}_{in}(\mathbf{x}_i)$ as

$$M_{in}(\mathbf{x}_i) = \frac{1}{n}\sum_{j=1}^n \left\{ 2\rho_n^{1/2}\mathbf{x}_i^T\mathbf{x}_{0j} - \frac{\mathbf{x}_i^T\mathbf{x}_{0j}\mathbf{x}_{0j}^T\mathbf{x}_i}{2\mathbf{x}_{0i}^T\mathbf{x}_{0j}} + (1 - \rho_n\mathbf{x}_{0i}^T\mathbf{x}_{0j})\log(1 - \rho_n^{1/2}\mathbf{x}_i^T\mathbf{x}_{0j}) \right\}.$$

Simple calculation shows that

$$\frac{\partial M_{in}}{\partial \mathbf{x}_i}(\mathbf{x}_i) = \frac{1}{n}\sum_{j=1}^n \rho_n^{1/2}\mathbf{x}_{0j}^T \left(\rho_n^{1/2}\mathbf{x}_{0i} - \mathbf{x}_i\right) \left\{ \frac{1}{\rho_n\mathbf{x}_{0i}^T\mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{1/2}\mathbf{x}_i^T\mathbf{x}_{0j}} \right\} \rho_n^{1/2}\mathbf{x}_{0j},$$

31

$$\frac{\partial^2 M_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\mathbf{x}_i) = -\frac{1}{n} \sum_{j=1}^n \left\{ \frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} + \frac{1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}}{(1 - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j})^2} \right\} \rho_n \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}},$$

and

$$\frac{\partial \widetilde{M}_{in}}{\partial \mathbf{x}_i}(\mathbf{x}_i) = \frac{1}{n} \sum_{j=1}^n \left(A_{ij} - \mathbf{x}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j \right) \left\{ \frac{1}{\widetilde{p}_{ij}} + \frac{1}{1 - \mathbf{x}_i \widetilde{\mathbf{x}}_j} \right\} \widetilde{\mathbf{x}}_j,$$

$$\frac{\partial^2 \widetilde{M}_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\mathbf{x}_i) = -\frac{1}{n} \sum_{j=1}^n \left\{ \frac{1}{\widetilde{p}_{ij}} + \frac{1 - A_{ij}}{(1 - \mathbf{x}_i \widetilde{\mathbf{x}}_j)^2} \right\} \widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}}.$$

For simplicity of notation, in what follows the large probability bounds are with regard to $n \geq N_{c,\delta,\lambda}$ for some large constant $N_{c,\delta,\lambda}$ depending on c, δ, λ.

Claim I (identifiability): For all $\epsilon > 0$,

$$\inf_{\|\mathbf{x}_i - \rho_n^{1/2} \mathbf{x}_{0i}\|_2 \geq \epsilon} \left\| \frac{\partial M_{in}}{\partial \mathbf{x}_i}(\mathbf{x}_i) \right\|_2 \geq \lambda \epsilon > \left\| \frac{\partial M_{in}}{\partial \mathbf{x}_i}(\rho_n^{1/2} \mathbf{x}_{0i}) \right\|_2 = 0.$$

Claim II (uniform convergence): With probability at least $1 - n^{-c}$,

$$\max_{i \in [n]} \sup_{\|\mathbf{x}_i\|_2 \leq 1} \left\| \mathbf{W}^{\mathrm{T}} \frac{\partial \widetilde{M}_{in}}{\partial \mathbf{x}_i}(\mathbf{W}\mathbf{x}_i) - \frac{\partial M_{in}}{\partial \mathbf{x}_i}(\mathbf{x}_i) \right\|_2 \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n \rho_n}}.$$

Now we show Claim I. It is obvious that $\frac{\partial M_{in}}{\partial \mathbf{x}_i}(\rho_n^{1/2} \mathbf{x}_{0i}) = \mathbf{0}_d$. Because $\rho_n \leq 1$, $\|\mathbf{x}_i\|_2 \leq 1$, and $\max_{j \in [n]} \|\mathbf{x}_{0j}\|_2 \leq 1$, we have

$$-\frac{\partial^2 M_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\mathbf{x}_i) \succeq \frac{1}{n} \sum_{j=1}^n \frac{\mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}}}{\mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} \succeq \frac{1}{n} \sum_{j=1}^n \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}} \succeq \frac{1}{n} \mathbf{X}_0^{\mathrm{T}} \mathbf{X} \succeq \lambda_d \left(\frac{1}{n} \mathbf{X}_0^{\mathrm{T}} \mathbf{X} \right) \mathbf{I}_d \succeq \lambda \mathbf{I}_d,$$

which implies that $M_{in}(\mathbf{x}_i)$ is strictly concave with $\rho_n^{1/2} \mathbf{x}_{0i}$ as a unique maximizer. By Taylor's theorem, $\frac{\partial M_{in}}{\partial \mathbf{x}_i}(\mathbf{x}_i) = \frac{\partial^2 M_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i)(\mathbf{x}_i - \rho_n^{1/2} \mathbf{x}_{0i})$, where $\bar{\mathbf{x}}_i = \theta \rho_n^{1/2} \mathbf{x}_{0i} + (1 - \theta) \mathbf{x}_i$ for some $\theta \in [0, 1]$. It follows that

$$\left\| \frac{\partial M_{in}}{\partial \mathbf{x}_i}(\mathbf{x}_i) \right\|_2 = \left\| \frac{\partial^2 M_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i) \left(\mathbf{x}_i - \rho_n^{1/2} \mathbf{x}_{0i} \right) \right\|_2$$

$$\geq \lambda_d \left(-\frac{\partial^2 M_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i) \right) \left\| \mathbf{x}_i - \rho_n^{1/2} \mathbf{x}_{0i} \right\|_2 \geq \lambda \left\| \mathbf{x}_i - \rho_n^{1/2} \mathbf{x}_{0i} \right\|_2,$$

so $\inf_{\|\mathbf{x}_i - \rho_n^{1/2} \mathbf{x}_{0i}\|_2 \geq \epsilon} \left\| \frac{\partial M_{in}}{\partial \mathbf{x}_i}(\mathbf{x}_i) \right\|_2 \geq \lambda \epsilon$. Thus Claim I is shown. Now we show Claim II. By triangle inequality,

$$\left\| \mathbf{W}^{\mathrm{T}} \frac{\partial \widetilde{M}_{in}}{\partial \mathbf{x}_i}(\mathbf{W}^{\mathrm{T}} \mathbf{x}_i) - \frac{\partial M_{in}}{\partial \mathbf{x}_i}(\mathbf{x}_i) \right\|_2$$

$$\leq \left\| \frac{1}{n} \sum_{j=1}^{n} \left(A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j} \right) \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j}} \right) \rho_n^{1/2} \mathbf{x}_{0j} \right\|_2$$

$$+ \left\| \frac{1}{n} \sum_{j=1}^{n} \left(\mathbf{x}_i^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j} \right) \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j}} \right) \rho_n^{1/2} \mathbf{x}_{0j} \right\|_2$$

$$+ \left\| \frac{1}{n} \sum_{j=1}^{n} \left(A_{ij} - \mathbf{x}_i^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \right) \left\{ \left(\frac{1}{\widetilde{p}_{ij}} + \frac{1}{1 - \mathbf{x}_i^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j} \right) - \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{\frac{1}{2}} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j}} \right) \right\} \widetilde{\mathbf{x}}_j \right\|_2$$

$$+ \left\| \frac{1}{n} \sum_{j=1}^{n} \left(A_{ij} - \mathbf{x}_i^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \right) \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{\frac{1}{2}} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j}} \right) \left(\mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j - \rho_n^{\frac{1}{2}} \mathbf{x}_{0j} \right) \right\|_2 .$$

For the second term,

$$\max_{i \in [n]} \sup_{\|\mathbf{x}_i\|_2 \leq 1} \left\| \frac{1}{n} \sum_{j=1}^{n} \left(\mathbf{x}_i^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j} \right) \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j}} \right) \rho_n^{1/2} \mathbf{x}_{0j} \right\|_2$$

$$\leq \max_{i \in [n]} \sup_{\|\mathbf{x}_i\|_2 \leq 1} \frac{1}{n} \sum_{j=1}^{n} \left\| \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j - \rho_n^{\frac{1}{2}} \mathbf{x}_{0j} \right\|_2 \|\mathbf{x}_i\|_2 \left(\left| \frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} \right| + \left| \frac{1}{1 - \rho_n^{1/2} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j}} \right| \right) \rho_n^{1/2} \|\mathbf{x}_{0j}\|_2$$

$$\lesssim_\delta \left\| \widetilde{\mathbf{X}} \mathbf{W} - \rho_n^{1/2} \mathbf{X}_0 \right\|_{2 \to \infty} \rho_n^{-1/2}$$

$$\lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n \rho_n}}$$

with probability at least $1 - n^{-c}$. For the third term,

$$\max_{i \in [n]} \sup_{\|\mathbf{x}_i\|_2 \leq 1} \left\| \frac{1}{n} \sum_{j=1}^{n} \left(A_{ij} - \mathbf{x}_i^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j \right) \right.$$

$$\times \left. \left\{ \left(\frac{1}{\widetilde{p}_{ij}} + \frac{1}{1 - \mathbf{x}_i^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j} \right) - \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{\frac{1}{2}} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j}} \right) \right\} \widetilde{\mathbf{x}}_j \right\|_2$$

$$\leq \max_{i \in [n]} \sup_{\|\mathbf{x}_i\|_2 \leq 1} \frac{1}{n} \sum_{j=1}^{n} \left(A_{ij} + \|\mathbf{x}_i\|_2 \|\widetilde{\mathbf{x}}_j\|_2 \right)$$

$$\times \left(\frac{|\widetilde{p}_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}|}{|\widetilde{p}_{ij} \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}|} + \frac{\left\| \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j - \rho_n^{\frac{1}{2}} \mathbf{x}_{0j} \right\|_2 \|\mathbf{x}_i\|_2}{(1 - \rho_n^{\frac{1}{2}} \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{0j})(1 - \mathbf{x}_i^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \widetilde{\mathbf{x}}_j)} \right) \|\widetilde{\mathbf{x}}_j\|_2$$

$$\lesssim_{c,\delta,\lambda} \max_{i \in [n]} \frac{1}{n} \sum_{j=1}^{n} (A_{ij} + \rho_n^{1/2}) \left(\rho_n^{-3/2} \sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log n}{n}} \right) \rho_n^{1/2}$$

$$\lesssim_{c,\delta,\lambda} \left(\frac{1}{n} \|\mathbf{A}\|_\infty + \rho_n^{1/2} \right) \rho_n^{-1} \sqrt{\frac{\log n}{n}}$$

$$\lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}$$

with probability at least $1 - n^{-c}$, where the second inequality follows from Lemma 14, and the last one from $\|\mathbf{A}\|_\infty \lesssim_c n\rho_n$ with probability at least $1 - n^{-c}$, which follows from Bernstein's inequality and triangle inequality.

For the fourth term,

$$\max_{i\in[n]} \sup_{\|\mathbf{x}_i\|_2 \le 1} \left\| \frac{1}{n} \sum_{j=1}^n \left(A_{ij} - \mathbf{x}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\right) \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{\frac{1}{2}}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_{0j}}\right) \left(\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \rho_n^{\frac{1}{2}}\mathbf{x}_{0j}\right) \right\|_2$$

$$\le \max_{i\in[n]} \sup_{\|\mathbf{x}_i\|_2 \le 1} \frac{1}{n} \sum_{j=1}^n (A_{ij} + \|\mathbf{x}_i\|_2\|\widetilde{\mathbf{x}}_j\|_2) \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{\frac{1}{2}}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_{0j}}\right) \left\|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \rho_n^{\frac{1}{2}}\mathbf{x}_{0j}\right\|_2$$

$$\lesssim_{c,\delta,\lambda} \max_{i\in[n]} \frac{1}{n} \sum_{j=1}^n (A_{ij} + \rho_n^{1/2})\rho_n^{-1} \left\|\widetilde{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\right\|_{2\to\infty}$$

$$\lesssim_{c,\delta,\lambda} \left(\frac{1}{n}\|\mathbf{A}\|_\infty + \rho_n^{1/2}\right)\rho_n^{-1}\sqrt{\frac{\log n}{n}} \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}$$

with probability at least $1 - n^{-c}$.

In order to bound the first term, a maximal inequality is required. We use the results in Chapter 8 of Kosorok (2008). Define a stochastic process on $\{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\|_2 \le 1\}$ for each $k \in [d]$,

$$J_{ink}(\mathbf{y}) = \frac{1}{n} \sum_{j=1}^n \left(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}\right) \left(\frac{1}{\rho_n \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{1/2}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_{0j}}\right) \rho_n^{1/2} x_{0jk}.$$

Then for all $\mathbf{y}, \mathbf{y}'$ with $\|\mathbf{y}\|_2 \le 1, \|\mathbf{y}'\|_2 \le 1$,

$$\left|J_{ink}(\mathbf{y}) - J_{ink}(\mathbf{y}')\right| = \left|\frac{1}{n} \sum_{j=1}^n \left(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}\right) \frac{\rho_n^{1/2}\mathbf{x}_{0j}^{\mathrm{T}}(\mathbf{y} - \mathbf{y}')}{(1 - \rho_n^{1/2}\overline{\mathbf{y}}^{\mathrm{T}}\mathbf{x}_{0j})^2} \rho_n^{1/2} x_{0jk}\right|,$$

where $\overline{\mathbf{y}} = \theta\mathbf{y} + (1-\theta)\mathbf{y}'$ for some $\theta \in [0,1]$. By Hoeffding's inequality,

$$\mathbb{P}\{\left|J_{ink}(\mathbf{y}) - J_{ink}(\mathbf{y}')\right| \ge t\} \le 2\exp\left\{-\frac{2n^2 t^2}{\sum_{j=1}^n (\rho_n^{1/2}\mathbf{x}_{0j}^{\mathrm{T}}(\mathbf{y} - \mathbf{y}'))^2 \rho_n x_{0jk}^2 / (1 - \rho_n^{1/2}\overline{\mathbf{y}}^{\mathrm{T}}\mathbf{x}_{0j})^4}\right\}$$

$$\le 2\exp\left\{-\frac{nt^2}{C_\delta}\rho_n^2\|\mathbf{y} - \mathbf{y}'\|_2^2\right\},$$

where $C_\delta > 0$ is a constant depending on δ, which indicates that $J_{ink}(\mathbf{y})$ is a sub-Gaussian process on $\{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\|_2 \le 1\}$ with respect to the metric $d_n(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|_2 \sqrt{C_\delta}\rho_n^2/n$. The metric entropy of the metric space $(\{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\|_2 \le 1\}, d_n)$ can be bounded by

$$\log D(\epsilon, \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\|_2 \le 1\}, d_n) \le d\log\left(\frac{K_\delta}{\epsilon}\sqrt{\frac{\rho_n^2}{n}}\right),$$

where K_δ is a constant depending on δ. Recall that the ψ_2-Orlicz norm (sub-Gaussian norm) of a random variable X is defined as

$$\|X\|_{\psi_2} = \inf\left\{c > 0 : \mathbb{E}\psi_2\left(\frac{X}{c}\right) \leq 1\right\},$$

where $\psi_2(x) = e^{x^2} - 1$ (see Chapter 8 of Kosorok, 2008).

By Theorem 8.4 in Kosorok (2008),

$$\left\|\sup_{\|\mathbf{y}\|_2 \leq 1} J_{ink}(\mathbf{y})\right\|_{\psi_2} \lesssim \int_0^{2\sqrt{\frac{4\rho_n^2}{n\delta^4}}} \sqrt{\log D(\epsilon, \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\|_2 \leq 1\}, d_n)} \mathrm{d}\epsilon$$

$$\leq \int_0^{2\sqrt{\frac{4\rho_n^2}{n\delta^4}}} \sqrt{d\log\left(\frac{K_\delta}{\epsilon}\sqrt{\frac{\rho_n^2}{n}}\right)} \mathrm{d}\epsilon$$

$$= \int_{K_\delta}^\infty K_\delta\sqrt{d}\sqrt{\frac{\rho_n^2}{n}}\sqrt{u}e^{-u}\mathrm{d}u \lesssim_{\delta,\lambda} \sqrt{\frac{\rho_n^2}{n}},$$

where we note that d depends on λ implicitly. Then by Lemma 8.1 in Kosorok (2008) and a union bound over $i \in [n]$, $\max_{i\in[n]}\sup_{\|\mathbf{x}_i\|_2\leq 1}|J_{ink}(\mathbf{x}_i)| \lesssim_{c,\delta,\lambda} \sqrt{(\rho_n^2\log n)/n}$ with probability at least $1 - n^{-c}$. So

$$\max_{i\in[n]}\sup_{\|\mathbf{x}_i\|_2\leq 1}\left\|\frac{1}{n}\sum_{j=1}^n\left(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}\right)\left(\frac{1}{\rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}} + \frac{1}{1 - \rho_n^{1/2}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_{0j}}\right)\rho_n^{1/2}\mathbf{x}_{0j}\right\|_2$$

$$\leq \max_{i\in[n]}\sup_{\|\mathbf{x}_i\|_2\leq 1}\sum_{k=1}^d|J_{ink}(\mathbf{x}_i)| \lesssim_{c,\delta,\lambda} \rho_n\sqrt{\frac{\log n}{n}}$$

with probability at least $1 - n^{-c}$. Thus Claim II is shown.

By Theorem 3, $\widehat{\mathbf{x}}_i$ is the unique zero of $\|\partial\widetilde{M}_{in}/\partial\mathbf{x}_i(\mathbf{x}_i)\|_2$ with probability at least $1 - n^{-c}$. Now

$$\max_{i\in[n]}\left(\left\|\frac{\partial M_{in}}{\partial\mathbf{x}_i}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i)\right\|_2 - \left\|\frac{\partial M_{in}}{\partial\mathbf{x}_i}(\rho_n^{1/2}\mathbf{x}_{0i})\right\|_2\right)$$

$$\leq \max_{i\in[n]}\left(\left\|\frac{\partial M_{in}}{\partial\mathbf{x}_i}(\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i)\right\|_2 - \left\|\frac{\partial\widetilde{M}_{in}}{\partial\mathbf{x}_i}(\widehat{\mathbf{x}}_i)\right\|_2\right)$$

$$+ \max_{i\in[n]}\left(\left\|\frac{\partial\widetilde{M}_{in}}{\partial\mathbf{x}_i}(\mathbf{W}\rho_n^{1/2}\mathbf{x}_{0i})\right\|_2 - \left\|\frac{\partial M_{in}}{\partial\mathbf{x}_i}(\rho_n^{1/2}\mathbf{x}_{0i})\right\|_2\right)$$

$$\leq 2\max_{i\in[n]}\sup_{\|\mathbf{x}_i\|_2\leq 1}\left\|\mathbf{W}^{\mathrm{T}}\frac{\partial\widetilde{M}_{in}}{\partial\mathbf{x}_i}(\mathbf{W}\mathbf{x}_i) - \frac{\partial M_{in}}{\partial\mathbf{x}_i}(\mathbf{x}_i)\right\|_2 \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}},$$

where the first inequality follows from $\widehat{\mathbf{x}}_i$ being the unique zero of $\|\partial\widetilde{M}_{in}/\partial\mathbf{x}_i(\mathbf{x}_i)\|_2$ with probability at least $1 - n^{-c}$, the second inequality from triangle inequality, and the third inequality from Claim II.

35

By Claim I, take $\epsilon = K_{c,\delta,\lambda}\sqrt{(\log n)/(n\rho_n)}$, we have

$$\max_{i\in[n]}\left\|\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}\right\|_2 \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}$$

with probability at least $1 - n^{-c}$.

■ We next establish the asymptotic normality. We utilize the asymptotic normality of the one-step estimator $\widehat{\mathbf{x}}_i^{(\mathrm{OS})}$ (Theorem 18) to establish the asymptotic normality of the maximum surrogate likelihood estimator $\widehat{\mathbf{x}}_i$. By the previous part of the theorem, we know that with probability at least $1 - n^{-c}$, $\widehat{\mathbf{x}}_i$ is in the interior of the closed unit ball $B(\mathbf{0}_d, 1) = \{\mathbf{x}\in\mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$. For each $k\in[d]$, we apply Taylor's theorem to $(\partial\widetilde{M}_{in})/(\partial x_{ik})(\widehat{\mathbf{x}}_i) = 0$ at $\mathbf{x}_i = \widetilde{\mathbf{x}}_i$ to obtain

$$0 = \frac{\partial\widetilde{M}_{in}}{\partial x_{ik}}(\widehat{\mathbf{x}}_i) = \frac{\partial\widetilde{M}_{in}}{\partial x_{ik}}(\widetilde{\mathbf{x}}_i) + \frac{\partial}{\partial\widetilde{\mathbf{x}}_i^{\mathrm{T}}}\frac{\partial\widetilde{M}_{in}}{\partial x_{ik}}(\widetilde{\mathbf{x}}_i)(\widehat{\mathbf{x}}_i - \widetilde{\mathbf{x}}_i)$$
$$+ \frac{1}{2}(\widehat{\mathbf{x}}_i - \widetilde{\mathbf{x}}_i)^{\mathrm{T}}\frac{\partial^2}{\partial\mathbf{x}_i\partial\mathbf{x}_i^{\mathrm{T}}}\frac{\partial\widetilde{M}_{in}}{\partial x_{ik}}(\bar{\mathbf{x}}_i)(\widehat{\mathbf{x}}_i - \widetilde{\mathbf{x}}_i),$$

where $\bar{\mathbf{x}}_i = \theta\widehat{\mathbf{x}}_i + (1 - \theta)\widetilde{\mathbf{x}}_i$ for some $\theta\in[0, 1]$. It is easy to compute

$$\frac{\partial^2}{\partial\mathbf{x}_i\partial\mathbf{x}_i^{\mathrm{T}}}\frac{\partial\widetilde{M}_{in}}{\partial x_{ik}}(\mathbf{x}_i) = -\frac{2}{n}\sum_{j=1}^n\frac{(1 - A_{ij})\widetilde{x}_{jk}}{(1 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)^3}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}},$$

then

$$\sup_{\|\mathbf{x}_i\|_2\leq 1}\left\|\frac{\partial^2}{\partial\mathbf{x}_i\partial\mathbf{x}_i^{\mathrm{T}}}\frac{\partial\widetilde{M}_{in}}{\partial x_{ik}}(\mathbf{x}_i)\right\|_2 = \sup_{\|\mathbf{x}_i\|_2\leq 1}\left\|\frac{2}{n}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathrm{diag}\left\{\frac{1 - A_{i1}}{(1 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_1)^3}, \ldots, \frac{1 - A_{in}}{(1 - \mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_n)^3}\right\}\widetilde{\mathbf{X}}\right\|_2$$
$$\lesssim_\delta \|\frac{1}{n}\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}\|_2 = \|\frac{1}{n}\mathbf{A}\|_2 \leq \frac{1}{n}\left(\|\mathbf{A} - \mathbf{P}\|_2 + \|\mathbf{P}\|_2\right) \lesssim_c \rho_n,$$

where in the last inequality we applied the fact that $\|\mathbf{A} - \mathbf{P}\|_2 \lesssim_c \sqrt{n\rho_n}$ with probability at least $1 - n^{-c}$ (Theorem 5.2 in Lei and Rinaldo, 2015). By Lemma 13 and the previous part of the theorem, with probability at least $1 - n^{-c}$,

$$\|\widehat{\mathbf{x}}_i - \widetilde{\mathbf{x}}_i\|_2 \leq \|\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 + \|\mathbf{W}^{\mathrm{T}}\widetilde{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}.$$

So the Taylor expansion of $(\partial\widetilde{M}_{in})/(\partial\mathbf{x}_i)$ mentioned above can be written as

$$-\left(\frac{\partial^2\widetilde{M}_{in}}{\partial\mathbf{x}_i\partial\mathbf{x}_i^{\mathrm{T}}}(\widetilde{\mathbf{x}}_i) + \mathbf{R}_{in1}\right)(\widehat{\mathbf{x}}_i - \widetilde{\mathbf{x}}_i) = \frac{\partial\widetilde{M}_{in}}{\partial\mathbf{x}_i}(\widetilde{\mathbf{x}}_i),$$

where $\mathbf{R}_{in1}\in\mathbb{R}^{d\times d}$ is a random matrix with $\|\mathbf{R}_{in1}\|_2 \lesssim_{c,\delta,\lambda} \rho_n^{1/2}\sqrt{(\log n)/n}$ with probability at least $1 - n^{-c}$. By definition of $\widetilde{M}_{in}(\mathbf{x}_i)$ and Lemma 15,

$$\left(\frac{1}{n}\sum_{j=1}^n\frac{1}{\widetilde{p}_{ij}(1 - \widetilde{p}_{ij})}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}} + \mathbf{R}_{in2}\right)(\widehat{\mathbf{x}}_i - \widetilde{\mathbf{x}}_i) = \frac{1}{n}\sum_{j=1}^n\frac{A_{ij} - \widetilde{p}_{ij}}{\widetilde{p}_{ij}(1 - \widetilde{p}_{ij})}\widetilde{\mathbf{x}}_j,$$

where $\mathbf{R}_{in2} \in \mathbb{R}^{d \times d}$ is a random matrix with $\|\mathbf{R}_{in2}\|_2 \lesssim_{c,\delta,\lambda} \rho_n^{1/2} \sqrt{(\log n)/n}$ with probability at least $1 - n^{-c}$ and $\widetilde{p}_{ij} = \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j$, $i, j \in [n]$.

Denote $\widetilde{\mathbf{G}}_{in} = \frac{1}{n} \sum_{j=1}^{n} \frac{\widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}}}{\widetilde{p}_{ij}(1 - \widetilde{p}_{ij})}$. Similarly as in the proof of Theorem 3,

$$\frac{\lambda}{2} \leq \frac{1}{n\rho_n} \lambda_d(\mathbf{A}) = \lambda_d \left(\frac{1}{n\rho_n} \widetilde{\mathbf{X}}^{\mathrm{T}} \widetilde{\mathbf{X}} \right) \leq \lambda_d(\widetilde{\mathbf{G}}_{in}) \leq \lambda_1(\widetilde{\mathbf{G}}_{in})$$

$$\lesssim_\delta \lambda_1 \left(\frac{1}{n\rho_n} \widetilde{\mathbf{X}}^{\mathrm{T}} \widetilde{\mathbf{X}} \right) = \frac{1}{n\rho_n} \lambda_1(\mathbf{A}) \lesssim_c 1,$$

that is, $\widetilde{\mathbf{G}}_{in}$ is finite and positive definite with probability at least $1 - n^{-c}$. Now write

$$\widehat{\mathbf{x}}_i - \widetilde{\mathbf{x}}_i = \left(\widetilde{\mathbf{G}}_{in} + \mathbf{R}_{in2} \right)^{-1} \frac{1}{n} \sum_{j=1}^{n} \frac{A_{ij} - \widetilde{p}_{ij}}{\widetilde{p}_{ij}(1 - \widetilde{p}_{ij})} \widetilde{\mathbf{x}}_j$$

$$= \left(\mathbf{I}_d + \widetilde{\mathbf{G}}_{in}^{-1} \mathbf{R}_{in2} \right)^{-1} \widetilde{\mathbf{G}}_{in}^{-1} \frac{1}{n} \sum_{j=1}^{n} \frac{A_{ij} - \widetilde{p}_{ij}}{\widetilde{p}_{ij}(1 - \widetilde{p}_{ij})} \widetilde{\mathbf{x}}_j$$

$$= \sum_{m=0}^{\infty} (-\widetilde{\mathbf{G}}_{in}^{-1} \mathbf{R}_{in2})^m (\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \widetilde{\mathbf{x}}_i)$$

$$= (\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \widetilde{\mathbf{x}}_i) + \sum_{m=1}^{\infty} (-\widetilde{\mathbf{G}}_{in}^{-1} \mathbf{R}_{in2})^m (\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \widetilde{\mathbf{x}}_i),$$

then

$$\|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_i^{(\mathrm{OS})}\|_2 \leq \sum_{m=1}^{\infty} \|\widetilde{\mathbf{G}}_{in}^{-1}\|_2^m \|\mathbf{R}_{in2}\|_2^m \|\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \widetilde{\mathbf{x}}_i\|_2$$

$$= \frac{\|\widetilde{\mathbf{G}}_{in}^{-1}\|_2 \|\mathbf{R}_{in2}\|_2}{1 - \|\widetilde{\mathbf{G}}_{in}^{-1}\|_2 \|\mathbf{R}_{in2}\|_2} \|\widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \widetilde{\mathbf{x}}_i\|_2$$

$$\lesssim_{c,\delta\lambda} \rho_n^{1/2} \sqrt{\frac{\log n}{n}} \left(\|\mathbf{W}^{\mathrm{T}} \widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \rho_n^{1/2} \mathbf{x}_{0i}\|_2 + \|\widetilde{\mathbf{X}} \mathbf{W} - \rho_n^{1/2} \mathbf{X}_0\|_{2 \to \infty} \right).$$

Note that

$$\lambda \leq \lambda_d \left(\frac{1}{n} \mathbf{X}_0^{\mathrm{T}} \mathbf{X}_0 \right) \leq \lambda_d(\mathbf{G}_{0in}) \leq \lambda_1(\mathbf{G}_{0in}) \leq \frac{1}{\delta^2} \lambda_1 \left(\frac{1}{n} \mathbf{X}_0^{\mathrm{T}} \mathbf{X}_0 \right) \leq \frac{1}{\delta^2}, \tag{16}$$

that is, $\mathbf{G}_{0in}$ is positive definite with eigenvalues bounded away from 0 and ∞. By Theorem 18 and Bernstein's inequality, $\|\mathbf{W}^{\mathrm{T}} \widehat{\mathbf{x}}_i^{(\mathrm{OS})} - \rho_n^{1/2} \mathbf{x}_{0i}\|_2 \lesssim_{c,\delta,\lambda} \sqrt{\log n/n}$ with probability at least $1 - (n\rho_n)^{-c}$. By Lemma 13, $\|\widetilde{\mathbf{X}} \mathbf{W} - \rho_n^{1/2} \mathbf{X}_0\|_{2 \to \infty} \lesssim_{c,\delta,\lambda} \sqrt{\log n/n}$ with probability at least $1 - n^{-c}$. So $\|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_i^{(\mathrm{OS})}\|_2 \lesssim_{c,\delta,\lambda} \rho_n^{1/2} (\log n)/n$ with probability at least $1 - (n\rho_n)^{-c}$. By Theorem 18 and Slutsky's theorem, we have

$$\sqrt{n} \mathbf{G}_{0in}^{1/2} \left(\mathbf{W}^{\mathrm{T}} \widehat{\mathbf{x}}_i - \rho_n^{1/2} \mathbf{x}_{0i} \right) \xrightarrow{\mathcal{L}} N(\mathbf{0}_d, \mathbf{I}_d),$$

37

and

$$\mathbf{G}_{0in}^{1/2}(\mathbf{W}^{\mathrm{T}}\hat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}) = \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{G}_{0in}^{-1/2}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} + \mathbf{r}_{in},$$

where

$$\|\mathbf{r}_{in}\|_2 \lesssim_{c,\delta,\lambda} \rho_n^{1/2}\frac{\log n}{n} + \frac{1}{\sqrt{n}}\sqrt{\frac{(\log(n\rho_n))^4}{n\rho_n}}$$

with probability at least $1 - (n\rho_n)^{-c}$.

■ We finally show the convergence of the sum of squares errors, that is

$$\|\widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}(\mathbf{G}_{0in}^{-1}) \xrightarrow{\mathbb{P}_0} 0.$$

By the previous result, we have

$$\|\widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \left\| \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} + \mathbf{G}_{0in}^{-1/2}\mathbf{r}_{in} \right\|_2^2$$

$$= \sum_{i=1}^{n} \left\| \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} \right\|_2^2 + \sum_{i=1}^{n} \|\mathbf{G}_{0in}^{-1/2}\mathbf{r}_{in}\|_2^2$$

$$+ 2\sum_{i=1}^{n} \left\langle \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}, \mathbf{G}_{0in}^{-1/2}\mathbf{r}_{in} \right\rangle.$$

By Lemma 16, the first term equals

$$\sum_{i=1}^{n} \mathbb{E}_0 \left\| \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} \right\|_2^2 + o_{\mathbb{P}_0}(1)$$

$$= \frac{1}{n^2\rho_n} \sum_{i=1}^{n}\sum_{a=1}^{n}\sum_{b=1}^{n} \frac{\mathbb{E}_0\{(A_{ia} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0a})(A_{ib} - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0b})\}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0a}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0a})\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0b}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0b})}\mathbf{x}_{0a}^{\mathrm{T}}\mathbf{G}_{0in}^{-2}\mathbf{x}_{0b} + o_{\mathbb{P}_0}(1)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\mathbf{x}_{0j}^{\mathrm{T}}\mathbf{G}_{0in}^{-2}\mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})} + o_{\mathbb{P}_0}(1)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr}\left\{ \frac{1}{n}\sum_{j=1}^{n} \frac{\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}}{\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}(1 - \rho_n\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})}\mathbf{G}_{0in}^{-2} \right\} + o_{\mathbb{P}_0}(1)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr}(\mathbf{G}_{0in}^{-1}) + o_{\mathbb{P}_0}(1).$$

For the second term, by Theorem 4.7 in Xie (2024), we have

$$\sum_{i=1}^{n} \|\mathbf{r}_{in}\|_2^2 \le n \max_{i\in[n]} \|\mathbf{r}_{in}\|_2^2 \lesssim_{c,\delta,\lambda} \frac{\rho_n(\log n)^2}{n} + \frac{(\log n)^4}{n\rho_n}$$

with probability at least $1 - n^{-c}$ for all $n \geq N_{c,\delta,\lambda}$, so $\sum_{i=1}^{n} \|\mathbf{r}_{in}\|_2^2 = o_{\mathbb{P}_0}(1)$ by the condition that $(\log n)^4 = o(n\rho_n)$. For the third term, by Cauchy–Schwarz inequality, we have

$$
\left| \sum_{i=1}^{n} \left\langle \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}) \mathbf{G}_{0in}^{-1} \mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}(1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j})}, \mathbf{G}_{0in}^{-1/2} \mathbf{r}_{in} \right\rangle \right|
$$

$$
\leq \sum_{i=1}^{n} \left\| \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}) \mathbf{G}_{0in}^{-1} \mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}(1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j})} \right\|_2 \| \mathbf{G}_{0in}^{-1/2} \mathbf{r}_{in} \|_2
$$

$$
\leq \left\{ \sum_{i=1}^{n} \left\| \frac{1}{n\rho_n^{1/2}} \sum_{j=1}^{n} \frac{(A_{ij} - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}) \mathbf{G}_{0in}^{-1} \mathbf{x}_{0j}}{\mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}(1 - \rho_n \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j})} \right\|_2^2 \right\}^{1/2} \left\{ \sum_{i=1}^{n} \| \mathbf{G}_{0in}^{-1/2} \mathbf{r}_{in} \|_2^2 \right\}^{1/2}
$$

$$
= O_{\mathbb{P}_0}(1) \times o_{\mathbb{P}_0}(1) = o_{\mathbb{P}_0}(1).
$$

Hence, we conlcude that

$$
\| \widehat{\mathbf{X}} \mathbf{W} - \rho_n^{1/2} \mathbf{X}_0 \|_{\mathrm{F}}^2 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr}(\mathbf{G}_{0in}^{-1}) + o_{\mathbb{P}_0}(1).
$$

∎

B.3 Proof of Theorem 7

Proof For convenience, in this proof, we introduce the following $\widetilde{O}_{\mathbb{P}}(\cdot)$ notation. Given a sequence of random matrices $(\mathbf{X}_n)_{n=1}^{\infty}$ and a deterministic positive sequence $(\varepsilon_n)_{n=1}^{\infty}$, we write $\mathbf{X}_n = \widetilde{O}_{\mathbb{P}}(\varepsilon_n)$, if for all $c > 0$, there exists constants $C_c, N_c > 0$, such that $\mathbb{P}(\|\mathbf{X}_n\|_2 > C_c \varepsilon_n) \leq n^{-c}$ for all $n \geq N_c$. It is clear that $\widetilde{O}_{\mathbb{P}}(\cdot)$ is a stronger notion than $O_{\mathbb{P}}(\cdot)$. Let $\psi(s,t) : (0,1)^2 \to \mathbb{R}_+$ be a function such that $\psi(\mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}, \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}) = 1/\mathrm{var}(A_{ij})$. Denote by $\mathbf{g}_{ij}(\mathbf{x}, \mathbf{u}, \mathbf{v}) = (A_{ij} - \mathbf{x}^{\mathrm{T}} \mathbf{v}) \psi(\mathbf{x}^{\mathrm{T}} \mathbf{v}, \mathbf{u}^{\mathrm{T}} \mathbf{v})$. Consider a generic estimating equation $(1/n) \sum_{j=1}^{n} \mathbf{g}_{ij}(\mathbf{x}_i, \widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_j)$ (also known as the eigenvector-assisted estimating equation in Xie and Wu, 2024). A simple algebra shows that the solution to this estimating equation corresponds to the one-step estimator and the maximum surrogate likelihood estimator when $\psi(s,t) = 1/t + 1/(1-t)$ and $\psi(s,t) = 1/t + 1/(1-s)$, respectively. Therefore, it is sufficient to work with the generic estimating equation. With a slight abuse of notation, we denote by $\widehat{\mathbf{x}}_i$ the associated estimating equation estimator, and depending on the context, $\widehat{\mathbf{x}}_i$ may represent the one-step estimator or the maximum surrogate likelihood estimator in this subsection. Following the proof of Theorem 1 in Xie and Wu (2024), we have $\mathbf{W}^* \widehat{\mathbf{x}}_i - \mathbf{x}_{0i} = \boldsymbol{\gamma}_i + \widehat{\mathbf{r}}_i$, where $\boldsymbol{\gamma}_i = \mathbf{G}_{0in}^{-1}(1/n) \sum_{j=1}^{n} (A_{ij} - \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}) \mathbf{x}_{0j} / \{\mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j}(1 - \mathbf{x}_{0i}^{\mathrm{T}} \mathbf{x}_{0j})\}$, $\widehat{\mathbf{r}}_i = \widetilde{O}_{\mathbb{P}}\{(\log n)^{2\xi}/n\}$ for any $\xi > 1$, and $\mathbf{W}^*$ is a diagonal matrix whose kth diagonal entry is the sign of $\mathbf{u}_k^{\mathrm{T}} \widehat{\mathbf{u}}_k$, where $\widehat{\mathbf{u}}_k$ is the eigenvector of $\mathbf{A}$ corresponding to its kth largest eigenvalue $\widehat{\lambda}_k$. For the ASE, we have a similar first-order stochastic expansion $\mathbf{W}^* \widetilde{\mathbf{x}}_i - \mathbf{x}_{0i} = \mathbf{e}_i^{\mathrm{T}} \mathbf{E} \mathbf{X}_0 (\mathbf{X}_0^{\mathrm{T}} \mathbf{X}_0)^{-1} + \widetilde{\mathbf{r}}_i$, where $\widetilde{\mathbf{r}}_i = \widetilde{O}_{\mathbb{P}}\{(\log n)^{2\xi}/n\}$ for any $\xi > 1$. Furthermore, by Theorem 1 in Xie and Zhang (2025), the proof of Theorem 1 in Athreya et al. (2022), Lemma S2.3

39

of Xie (2024), and the equation

$$\widetilde{\mathbf{X}}\mathbf{W}^* - \mathbf{X}_0 = \mathbf{E}\mathbf{X}_0(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1} + (\mathbf{U}_{\mathbf{A}}\mathbf{W}^* - \mathbf{U}_{\mathbf{P}} - \mathbf{E}\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1})\mathbf{S}_{\mathbf{P}}^{1/2}$$
$$+ \mathbf{U}_{\mathbf{P}}(\mathbf{S}_{\mathbf{A}}^{1/2} - \mathbf{S}_{\mathbf{P}}^{1/2}) + (\mathbf{U}_{\mathbf{A}}\mathbf{W}^* - \mathbf{U}_{\mathbf{P}})(\mathbf{S}_{\mathbf{A}}^{1/2} + \mathbf{S}_{\mathbf{P}}^{1/2}),$$

we obtain the following second-order stochastic expansion of the ASE:

$$\mathbf{W}^*\widetilde{\mathbf{x}}_i - \mathbf{x}_{0i} = \widetilde{\boldsymbol{\alpha}}_i + \widetilde{\boldsymbol{\beta}}_i + \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}$$

for any $\xi > 1$, where $\widetilde{\boldsymbol{\alpha}}_i = [\widetilde{\alpha}_{i1}, \ldots, \widetilde{\alpha}_{id}]^{\mathrm{T}}$, and $\widetilde{\boldsymbol{\beta}}_i = [\widetilde{\beta}_{i1}, \ldots, \widetilde{\beta}_{id}]^{\mathrm{T}}$ are given by

$$\widetilde{\alpha}_{ik} = \mathbf{e}_i^{\mathrm{T}}\left(\mathbf{I}_n - \frac{1}{2}\mathbf{u}_k\mathbf{u}_k^{\mathrm{T}} - \sum_{m\in[d]\setminus\{k\}}\frac{\lambda_m\mathbf{u}_m\mathbf{u}_m^{\mathrm{T}}}{\lambda_k - \lambda_m}\right)\frac{\mathbf{E}\mathbf{u}_k}{\lambda_k^{1/2}},$$

$$\widetilde{\beta}_{ik} = \mathbf{e}_i^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{u}_k\mathbf{u}_k^{\mathrm{T}} - \sum_{m\in[d]\setminus\{k\}}\frac{\lambda_m\mathbf{u}_m\mathbf{u}_m^{\mathrm{T}}}{\lambda_k - \lambda_m}\right)\frac{\mathbf{E}^2\mathbf{u}_k}{\lambda_k^{3/2}}.$$

Here, $\mathbf{S}_{\mathbf{A}} = \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d)$, $\mathbf{S}_{\mathbf{P}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, $\mathbf{E} = [E_{ij}]_{n\times n} = \mathbf{A} - \mathbf{X}_0\mathbf{X}_0^{\mathrm{T}}$, $\mathbf{U}_{\mathbf{A}} = [\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_d]$, and $\mathbf{U}_{\mathbf{P}} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$. For any $a, b \in \{0, 1, 2\}$, denote by $D^{(a,b)}\psi(s,t) = \partial^{a+b}\psi(s,t)/(\partial^a s\partial^b t)$ and write $D^{(a,b)}\psi_{ij} = D^{(a,b)}\psi(\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}, \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})$. In particular, when $a = b = 0$, we write $D^{(0,0)}\psi_{ij} = \psi_{ij} = \psi(\mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j}, \mathbf{x}_{0i}^{\mathrm{T}}\mathbf{x}_{0j})$. Since $\mathbf{g}_{ij}(\cdot, \cdot, \cdot)$ is continuously three-times differentiable and $\|\widetilde{\mathbf{X}}\mathbf{W}^* - \mathbf{X}_0\|_{2\to\infty} + \|\widehat{\mathbf{X}}\mathbf{W}^* - \mathbf{X}_0\|_{2\to\infty} = \widetilde{O}_{\mathbb{P}}\{(\log n)^{3\xi}/\sqrt{n}\}$ for any $\xi > 1$, it follows Taylor's expansion that

$$\mathbf{G}_{0in}(\mathbf{W}^*\widehat{\mathbf{x}}_i - \mathbf{x}_{0i})$$

$$= \frac{1}{n}\sum_{j=1}^{n}\mathbf{g}_{ij}(\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j}) \tag{17}$$

$$+ \left\{\frac{1}{n}\sum_{j=1}^{n}\frac{\partial\mathbf{g}_{ij}}{\partial\mathbf{x}^{\mathrm{T}}}(\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j}) + \mathbf{G}_{0in}\right\}(\mathbf{W}^*\widehat{\mathbf{x}}_i - \mathbf{x}_{0i}) \tag{18}$$

$$+ \frac{1}{n}\sum_{j=1}^{n}\frac{\partial\mathbf{g}_{ij}}{\partial\mathbf{u}^{\mathrm{T}}}(\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j})(\mathbf{W}^*\widetilde{\mathbf{x}}_i - \mathbf{x}_{0i}) \tag{19}$$

$$+ \frac{1}{n}\sum_{j=1}^{n}\frac{\partial\mathbf{g}_{ij}}{\partial\mathbf{v}^{\mathrm{T}}}(\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j})(\mathbf{W}^*\widetilde{\mathbf{x}}_j - \mathbf{x}_{0j}) \tag{20}$$

$$+ \frac{1}{2n}\sum_{j=1}^{n}\sum_{k,l=1}^{d}\frac{\partial\mathbf{g}_{ij}}{\partial x_k\partial x_l}(\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j})(w_k^*\widehat{x}_{ik} - x_{0ik})(w_l^*\widehat{x}_{il} - x_{0il}) \tag{21}$$

$$+ \frac{1}{2n}\sum_{j=1}^{n}\sum_{k,l=1}^{d}\frac{\partial\mathbf{g}_{ij}}{\partial u_k\partial u_l}(\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j})(w_k^*\widetilde{x}_{ik} - x_{0ik})(w_l^*\widetilde{x}_{il} - x_{0il}) \tag{22}$$

$$+ \frac{1}{2n}\sum_{j=1}^{n}\sum_{k,l=1}^{d}\frac{\partial\mathbf{g}_{ij}}{\partial v_k\partial v_l}(\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j})(w_k^*\widetilde{x}_{jk} - x_{0jk})(w_l^*\widetilde{x}_{jl} - x_{0jl}) \tag{23}$$

$$+ \frac{1}{n} \sum_{j=1}^{n} \sum_{k,l=1}^{d} \frac{\partial \mathbf{g}_{ij}}{\partial x_k \partial u_l} (\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j}) (w_k^* \widehat{x}_{ik} - x_{0ik}) (w_l^* \widetilde{x}_{il} - x_{0il}) \tag{24}$$

$$+ \frac{1}{n} \sum_{j=1}^{n} \sum_{k,l=1}^{d} \frac{\partial \mathbf{g}_{ij}}{\partial x_k \partial v_l} (\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j}) (w_k^* \widehat{x}_{ik} - x_{0ik}) (w_l^* \widetilde{x}_{jl} - x_{0jl}) \tag{25}$$

$$+ \frac{1}{n} \sum_{j=1}^{n} \sum_{k,l=1}^{d} \frac{\partial \mathbf{g}_{ij}}{\partial u_k \partial v_l} (\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j}) (w_k^* \widetilde{x}_{ik} - x_{0ik}) (w_l^* \widetilde{x}_{jl} - x_{0jl}) \tag{26}$$

$$+ \widetilde{O}_{\mathbb{P}} \left\{ \frac{(\log n)^{3\xi}}{n^{3/2}} \right\},$$

where $\mathbf{W}^* = \mathrm{diag}(w_1^*, \ldots, w_d^*)$, $\mathbf{u} = [u_1, \ldots, u_d]^{\mathrm{T}}$, $\mathbf{v} = [v_1, \ldots, v_d]^{\mathrm{T}}$, $\widehat{\mathbf{x}}_i = [\widehat{x}_{i1}, \ldots, \widehat{x}_{id}]^{\mathrm{T}}$, and $\widetilde{\mathbf{x}}_i = [\widetilde{x}_{i1}, \ldots, \widetilde{x}_{id}]^{\mathrm{T}}$. Now we analyze each term separately. For the terms on (18) and (19), by Bernstein's inequality and the first-order expansions of $\widehat{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_i$, we have

$$\left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{\partial \mathbf{g}_{ij}}{\partial \mathbf{x}^{\mathrm{T}}} (\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j}) + \mathbf{G}_{0in} \right\} (\mathbf{W}^* \widehat{\mathbf{x}}_i - \mathbf{x}_{0i})$$

$$= \frac{1}{n} \sum_{j=1}^{n} E_{ij} D^{(1,0)} \psi_{ij} \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}} \boldsymbol{\gamma}_i + \widetilde{O}_{\mathbb{P}} \left\{ \frac{(\log n)^{3\xi}}{n^{3/2}} \right\},$$

$$\frac{1}{n} \sum_{j=1}^{n} \frac{\partial \mathbf{g}_{ij}}{\partial \mathbf{u}^{\mathrm{T}}} (\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j}) (\mathbf{W}^* \widetilde{\mathbf{x}}_i - \mathbf{x}_{0i})$$

$$= \frac{1}{n} \sum_{j=1}^{n} E_{ij} D^{(0,1)} \psi_{ij} \mathbf{x}_{0j} \mathbf{x}_{0j}^{\mathrm{T}} \mathbf{e}_i^{\mathrm{T}} \mathbf{E} \mathbf{X}_0 (\mathbf{X}_0^{\mathrm{T}} \mathbf{X}_0)^{-1/2} + \widetilde{O}_{\mathbb{P}} \left\{ \frac{(\log n)^{3\xi}}{n^{3/2}} \right\}$$

For the term on (20), by the second-order stochastic expansion for the ASE, Bernstein's inequality, Result S3 of Xie and Wu (2024), Result B.1 of Xie and Zhang (2025), and a union bound over $j \in \{1, \ldots, n\}$, we have

$$\frac{1}{n} \sum_{j=1}^{n} \frac{\partial \mathbf{g}_{ij}}{\partial \mathbf{v}^{\mathrm{T}}} (\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j}) (\mathbf{W}^* \widetilde{\mathbf{x}}_j - \mathbf{x}_{0j})$$

$$= \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{d} \frac{\partial \mathbf{g}_{ij}}{\partial v_k} (\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{y}_{0j}) (w_k^* \widetilde{x}_{jk} - x_{0jk})$$

$$= \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{d} E_{ij} \{ D^{(1,0)} \psi_{ij} \mathbf{x}_{0j} x_{0ik} + D^{(0,1)} \psi_{ij} \mathbf{x}_{0j} x_{0ik} + \psi_{ij} \mathbf{e}_k \} \widetilde{\alpha}_{jk}$$

$$+ \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{d} E_{ij} \{ D^{(1,0)} \psi_{ij} \mathbf{x}_{0j} x_{0ik} + D^{(0,1)} \psi_{ij} \mathbf{x}_{0j} x_{0ik} + \psi_{ij} \mathbf{e}_k \}$$

$$\times \mathbf{e}_j^{\mathrm{T}} \left(\mathbf{I}_n - \mathbf{u}_k \mathbf{u}_k - \sum_{m \in [d] \setminus \{k\}} \frac{\lambda_m \mathbf{u}_m \mathbf{u}_m^{\mathrm{T}}}{\lambda_k - \lambda_m} \right) \frac{\mathbf{E}^2 \mathbf{u}_k}{\lambda_k^{3/2}}$$

$$-\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\psi_{ij}\mathbf{x}_{0j}x_{0ik}(\widetilde{\alpha}_{jk}+\widetilde{\beta}_{jk})+\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}$$

$$=\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}E_{ij}\{D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}+D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}+\psi_{ij}\mathbf{e}_{k}\}\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{k}}{\lambda_{k}^{1/2}}$$

$$-\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\psi_{ij}\mathbf{x}_{0j}x_{0ik}(\widetilde{\alpha}_{jk}+\widetilde{\beta}_{jk})+\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}.$$

Next, we work with the second-order derivative terms. For the terms on (21) and (22), by the first-order expansion of $\widehat{\mathbf{x}}_i$ and the fact that $\|\widehat{\mathbf{X}}\mathbf{W}^*-\mathbf{X}_0\|_{2\to\infty}=\widetilde{O}_{\mathbb{P}}\{(\log n)/\sqrt{n}\}$, we have

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}\frac{\partial^2\mathbf{g}_{ij}}{\partial x_k\partial x_l}(\mathbf{x}_{0i},\mathbf{x}_{0i},\mathbf{x}_{0j})(w_l^*\widehat{x}_{il}-x_{0il})(w_k^*\widehat{x}_{ik}-x_{0ik})$$

$$=-\frac{2}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0jl}(w_l^*\widehat{x}_{il}-x_{0il})(w_k^*\widehat{x}_{ik}-x_{0ik})$$

$$+\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}E_{ij}D^{(2,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0jl}(w_l^*\widehat{x}_{il}-x_{0il})(w_k^*\widehat{x}_{ik}-x_{0ik})$$

$$=-\frac{2}{n}\sum_{j=1}^{n}D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\left[\boldsymbol{\gamma}_i+\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{2\xi}}{n}\right\}\right]\mathbf{x}_{0j}^{\mathrm{T}}\left[\boldsymbol{\gamma}_i+\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{2\xi}}{n}\right\}\right]$$

$$+\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}E_{ij}D^{(2,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0jl}\times\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{2\xi}}{n}\right\}$$

$$=-\frac{2}{n}\sum_{j=1}^{n}D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}(\mathbf{x}_{0j}^{\mathrm{T}}\boldsymbol{\gamma}_i)^2+\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}$$

and

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}\frac{\partial^2\mathbf{g}_{ij}}{\partial u_k\partial u_l}(\mathbf{x}_{0i},\mathbf{x}_{0i},\mathbf{x}_{0j})(w_l^*\widetilde{x}_{il}-x_{0il})(w_k^*\widetilde{x}_{ik}-x_{0ik})$$

$$=\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}E_{ij}D^{(0,2)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0jl}(w_l^*\widetilde{x}_{il}-x_{0il})(w_k^*\widetilde{x}_{ik}-x_{0ik})$$

$$=\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}E_{ij}D^{(0,2)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0jl}\times\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{2}}{n}\right\}=\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}.$$

For the term on (23), we first observe that, for any $(c_{ij}:i,j\in\{1,\ldots,n\})$ with $\sup_{i,j\in\{1,\ldots,n\}}|c_{ij}|=O(1)$, the following bound holds:

$$\frac{1}{n}\sum_{j=1}^{n}c_{ij}E_{ij}\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{k}}{\lambda_{k}^{1/2}}\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{l}}{\lambda_{l}^{1/2}}=\frac{c_{ii}E_{ii}}{n}\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{k}}{\lambda_{k}^{1/2}}\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{l}}{\lambda_{l}^{1/2}}+\frac{1}{n}\sum_{j\neq i}^{n}c_{ij}E_{ij}\sum_{a\neq i}^{n}\frac{E_{ja}x_{ak}}{\lambda_{k}}\sum_{b\neq i}^{n}\frac{E_{jb}x_{bl}}{\lambda_{l}}$$

$$+ \frac{1}{n}\sum_{j\neq i}^{n}\frac{c_{ij}x_{ik}E_{ij}^{2}}{\lambda_{k}}\sum_{b\neq i}\frac{E_{jb}x_{bl}}{\lambda_{l}} + \frac{1}{n}\sum_{j\neq i}^{n}\frac{c_{ij}x_{il}E_{ij}^{2}}{\lambda_{l}}\sum_{a\neq i}\frac{E_{ja}x_{ak}}{\lambda_{k}}$$

$$+ \frac{1}{n}\sum_{j\neq i}^{n}\frac{c_{ij}x_{ik}x_{il}E_{ij}^{3}}{\lambda_{k}\lambda_{l}} = \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}$$

for any $\xi > 1$, where we have used the Bernstein's inequality, the independence between $(E_{ij} : j \in \{1,\ldots,n\}\setminus\{i\})$ and $(E_{ja} : j,a \in \{1,\ldots,n\}\setminus\{i\})$, the fact that $|E_{ij}| \leq 1$ with probability one, and a union bound over $j \in \{1,\ldots,n\}$. Then, we write $\widetilde{\mathbf{r}}_i = [\widetilde{r}_{i1},\ldots,\widetilde{r}_{id}]^{\mathrm{T}}$ and obtain the following decomposition on the term on (23):

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}\frac{\partial^2\mathbf{g}_{ij}}{\partial v_k\partial v_l}(\mathbf{x}_{0i},\mathbf{x}_{0i},\mathbf{x}_{0j})(w_l^*\widetilde{x}_{jl} - x_{0jl})(w_k^*\widetilde{x}_{jk} - x_{0jk})$$

$$= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}E_{ij}\{D^{(2,0)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + 2D^{(1,1)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + D^{(0,2)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il}$$

$$+ D^{(1,0)}\psi_{ij}(\mathbf{e}_k x_{0il} + \mathbf{e}_l x_{0ik}) + D^{(0,1)}\psi_{ij}(\mathbf{e}_k x_{0il} + \mathbf{e}_l x_{0ik})\}$$

$$\times\left(\frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{Eu}_k}{\lambda_k^{1/2}} + \widetilde{r}_{jk}\right)\left(\frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{Eu}_l}{\lambda_l^{1/2}} + \widetilde{r}_{jl}\right)$$

$$- \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}\{2D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + 2D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + \psi_{ij}(\mathbf{e}_l x_{0ik} + \mathbf{e}_k x_{0il})\}$$

$$\times \frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{Eu}_k\mathbf{e}_j\mathbf{Eu}_l}{\lambda_k^{1/2}\lambda_l^{1/2}} + \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}$$

$$= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}E_{ij}\{D^{(2,0)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + 2D^{(1,1)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + D^{(0,2)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il}$$

$$+ D^{(1,0)}\psi_{ij}(\mathbf{e}_k x_{0il} + \mathbf{e}_l x_{0ik}) + D^{(0,1)}\psi_{ij}(\mathbf{e}_k x_{0il} + \mathbf{e}_l x_{0ik})\}\frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{Eu}_k}{\lambda_k^{1/2}}\frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{Eu}_l}{\lambda_l^{1/2}}$$

$$- \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}\{2D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + 2D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + \psi_{ij}(\mathbf{e}_l x_{0ik} + \mathbf{e}_k x_{0il})\}$$

$$\times \frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{Eu}_k\mathbf{e}_j\mathbf{Eu}_l}{\lambda_k^{1/2}\lambda_l^{1/2}} + \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}$$

$$= - \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}\{2D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + 2D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + \psi_{ij}(\mathbf{e}_l x_{0ik} + \mathbf{e}_k x_{0il})\}$$

$$\times \frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{Eu}_k\mathbf{e}_j\mathbf{Eu}_l}{\lambda_k^{1/2}\lambda_l^{1/2}} + \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\},$$

where we have used the first-order stochastic expansion for $\widetilde{\mathbf{x}}_i$ and a union bound over $j \in \{1,\ldots,n\}$. For the term on (24), by the first-order stochastic expansions of $\widehat{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_i$

43

and Bernstein's inequality, we obtain

$$
\frac{1}{n}\sum_{j=1}^{n}\sum_{k,l=1}^{d}\frac{\partial^2 \mathbf{g}_{ij}}{\partial x_k \partial u_l}(\mathbf{x}_{0i},\mathbf{x}_{0i},\mathbf{x}_{0j})(w_k^*\widehat{x}_{ik}-x_{0ik})(w_l^*\widetilde{x}_{il}-x_{0il})
$$

$$
=\frac{1}{n}\sum_{j=1}^{n}\sum_{k,l=1}^{d}E_{ij}D^{(1,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0jl}(w_k^*\widehat{x}_{ik}-x_{0ik})(w_l^*\widetilde{x}_{il}-x_{0il})
$$

$$
-\frac{1}{n}\sum_{j=1}^{n}\sum_{k,l=1}^{d}D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0jl}(w_k^*\widehat{x}_{ik}-x_{0ik})(w_l^*\widetilde{x}_{il}-x_{0il})
$$

$$
=\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{2\xi}}{n^{3/2}}\right\}-\frac{1}{n}\sum_{j=1}^{n}\sum_{k,l=1}^{d}D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0jl}\left[\gamma_{ik}\frac{\mathbf{e}_i^{\mathrm{T}}\mathbf{E}\mathbf{u}_l}{\lambda_l^{1/2}}+\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}\right]
$$

$$
=-\frac{1}{n}\sum_{j=1}^{n}D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\boldsymbol{\gamma}_i\mathbf{x}_{0j}^{\mathrm{T}}(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{E}\mathbf{e}_i+\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{2\xi}}{n^{3/2}}\right\}
$$

For the term on (25), by the first-order stochastic expansions of $\widehat{\mathbf{x}}_i$, $\widetilde{\mathbf{x}}_j$, Result 3 in Xie and Wu (2024), the fact that $|E_{ij}|\leq 1$ with probabiltiy one, and a union bound over $j\in\{1,\ldots,n\}$,

$$
\frac{1}{n}\sum_{j=1}^{n}\sum_{k,l=1}^{d}\frac{\partial^2 \mathbf{g}_{ij}}{\partial x_k \partial v_l}(\mathbf{x}_{0i},\mathbf{x}_{0i},\mathbf{y}_{0j})(w_k^*\widehat{x}_{ik}-x_{0ik})(w_l^*\widetilde{x}_{jl}-x_{0jl})
$$

$$
=\sum_{j=1}^{n}\sum_{k,l=1}^{d}\{D^{(2,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}+D^{(1,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}+D^{(1,0)}\psi_{ij}(\mathbf{x}_{0j}\mathbf{e}_k^{\mathrm{T}}\mathbf{e}_l+\mathbf{e}_l x_{0jk})\}
$$

$$
\times E_{ij}\left(\frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_l}{\lambda_l^{1/2}}+\widetilde{r}_{jl}\right)\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{\xi}}{n^{3/2}}\right\}
$$

$$
-\sum_{j=1}^{n}\sum_{k,l=1}^{d}\{2D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}+D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}+\psi_{ij}(\mathbf{e}_l x_{0jk}+\mathbf{x}_{0j}\mathbf{e}_k^{\mathrm{T}}\mathbf{e}_l)\}
$$

$$
\times\left(\frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_l}{\lambda_l^{1/2}}+\widetilde{r}_{jl}\right)\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{\xi}}{n^{3/2}}\right\}
$$

$$
=\sum_{j=1}^{n}\sum_{k,l=1}^{d}\{D^{(2,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}+D^{(1,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}+D^{(1,0)}\psi_{ij}(\mathbf{x}_{0j}\mathbf{e}_k^{\mathrm{T}}\mathbf{e}_l+\mathbf{e}_l x_{0jk})\}
$$

$$
\times\frac{E_{ij}\mathbf{e}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_l}{\lambda_l^{1/2}}\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{\xi}}{n^{3/2}}\right\}
$$

$$
+\sum_{j=1}^{n}\sum_{k,l=1}^{d}\{D^{(2,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}+D^{(1,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}+D^{(1,0)}\psi_{ij}(\mathbf{x}_{0j}\mathbf{e}_k^{\mathrm{T}}\mathbf{e}_l+\mathbf{e}_l x_{0jk})\}
$$

$$
\times E_{ij}\widetilde{r}_{jl}\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{\xi}}{n^{3/2}}\right\}
$$

$$- \sum_{j=1}^{n} \sum_{k=1}^{d} \sum_{l=1}^{d} \{2D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il} + D^{(0,1)}\psi_{ij}\mathbf{x}_{j}x_{0jk}x_{0il} + \psi_{ij}(\mathbf{e}_{l}x_{0jk} + \mathbf{x}_{0j}\mathbf{e}_{k}^{\mathrm{T}}\mathbf{e}_{l})\}$$

$$\times \frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{l}}{\lambda_{l}^{1/2}}\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{\xi}}{n^{3/2}}\right\}$$

$$- \sum_{j=1}^{n} \sum_{k=1}^{d} \sum_{l=1}^{d} \{2D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il} + D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il} + \psi_{ij}(\mathbf{e}_{l}x_{0jk} + \mathbf{x}_{0j}\mathbf{e}_{k}^{\mathrm{T}}\mathbf{e}_{l})\}$$

$$\times \widetilde{r}_{jl}\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{\xi}}{n^{3/2}}\right\}$$

$$= \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}.$$

Finally, for the term on (26), by the first-order stochastic expansions of $\widetilde{\mathbf{x}}_i$, $\widetilde{\mathbf{x}}_j$, Result 3 of Xie and Wu (2024), the fact that $|E_{ij}| \leq 1$ with probability one, Bernstein's inequality, and a union bound over $j \in \{1, \ldots, n\}$, we have

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{k,l=1}^{d} \frac{\partial^2 \mathbf{g}_{ij}}{\partial u_k \partial v_l}(\mathbf{x}_{0i}, \mathbf{x}_{0i}, \mathbf{x}_{0j})(w_k^* \widetilde{x}_{ik} - x_{0ik})(w_k^* \widetilde{x}_{jl} - x_{0jl})$$

$$= \sum_{j=1}^{n} \sum_{k,l=1}^{d} \{D^{(1,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il} + D^{(0,2)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il} + D^{(0,1)}\psi_{ij}(\mathbf{x}_{0j}\mathbf{e}_{k}^{\mathrm{T}}\mathbf{e}_{l} + \mathbf{e}_{l}x_{0jk})\}$$

$$\times E_{ij}\left(\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{l}}{\lambda_{l}^{1/2}} + \widetilde{r}_{jl}\right)\widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{\xi}}{n^{3/2}}\right\}$$

$$- \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{d} \sum_{l=1}^{d} D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}\left[\frac{\mathbf{e}_{i}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{k}}{\lambda_{k}^{1/2}} + \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{2\xi}}{n}\right\}\right]\left(\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{l}}{\lambda_{l}^{1/2}} + \widetilde{r}_{jl}\right)$$

$$= \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\} - \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{d} \sum_{l=1}^{d} D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0jk}x_{0il}\frac{\mathbf{e}_{i}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{k}}{\lambda_{k}^{1/2}}\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{l}}{\lambda_{l}^{1/2}} = \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\}.$$

Combining the above results, we obtain the following second-order expansion for $\widehat{\mathbf{x}}_i$:

$$\mathbf{W}^*\widehat{\mathbf{x}}_i - \mathbf{x}_{0i} = \boldsymbol{\gamma}_i - \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0i}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\widetilde{\boldsymbol{\alpha}}_j + \mathbf{q}_i + \widetilde{O}_{\mathbb{P}}\left\{\frac{(\log n)^{3\xi}}{n^{3/2}}\right\},$$

where

$$\mathbf{q}_i = \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}E_{ij}D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{E}\mathbf{e}_i + \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}E_{ij}D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\boldsymbol{\gamma}_i$$

$$+ \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}E_{ij}\{D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0ik} + D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0ik} + \psi_{ij}\mathbf{e}_k\}\frac{\mathbf{e}_{j}^{\mathrm{T}}\mathbf{E}\mathbf{u}_{k}}{\lambda_{k}^{1/2}}$$

$$- \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0i}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_j - \mathbf{G}_{in}^{-1}\frac{1}{n}\sum_{j=1}^{n}D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}(\mathbf{x}_{j}^{\mathrm{T}}\boldsymbol{\gamma}_i)^2$$

45

$$- \mathbf{G}_{0in}^{-1} \frac{1}{2n} \sum_{j=1}^{n} \sum_{k,l=1}^{d} \{2D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + 2D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}x_{0ik}x_{0il} + \psi_{ij}(\mathbf{e}_l x_{0ik} + \mathbf{e}_k x_{0il})\}$$

$$\times \frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{E}\mathbf{u}_k\mathbf{e}_j\mathbf{E}\mathbf{u}_l}{\lambda_k^{1/2}\lambda_l^{1/2}} - \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}D^{(0,1)}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\gamma_i\mathbf{x}_{0j}^{\mathrm{T}}(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{E}\mathbf{e}_i.$$

Since $|E_{ij}| \le 1$ with probability one, it follows that $\mathbb{E}\mathbf{W}^*\widehat{\mathbf{x}}_i = \mathbf{x}_{0i} + \mathbb{E}\mathbf{q}_i + o(n^{-1})$. Furthermore, a simple algebra shows that

$$\mathbb{E}\mathbf{q}_i = \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}D^{(0,1)}\psi_{ij}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{x}_{0j} + \mathbf{G}_{0in}^{-1}\frac{1}{n}\sum_{j=1}^{n}D^{(1,0)}\psi_{ij}\mathbf{x}_{0j}\mathbf{x}_{0j}^{\mathrm{T}}\mathbf{G}_{0in}^{-1}\mathbf{x}_{0j}$$

$$+ \mathbf{b}_i^{(\mathrm{ASE})} + \mathbf{b}_i^{(\mathrm{base})}.$$

The proof is completed by substituting the generic function $\psi(s,t)$ above with $\psi(s,t) = 1/t + 1/(1-t)$ for the one-step estimator and $\psi(s,t) = 1/t + 1/(1-s)$ for the maximum surrogate likelihood estimator, respectively. ∎

B.4 Proof of Theorem 11

Proof Similar to the earlier proofs, the large probability bounds below are with regard to $n \ge N_{c,\delta,\lambda}$ for some large constant $N_{c,\delta,\lambda}$ depending on c, δ, λ. By definition, $\mathbf{t} = \sqrt{n}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_i - \widehat{\mathbf{x}}_i)$, then $\mathbf{x}_i = \widehat{\mathbf{x}}_i + \mathbf{W}\mathbf{t}/\sqrt{n}$. Denote the parameter space of $\mathbf{t}$ by $\widehat{\Theta}_{in} = \{\mathbf{t} \in \mathbb{R}^d : \|\widehat{\mathbf{x}}_i + \mathbf{W}\mathbf{t}/\sqrt{n}\|_2 \le 1\}$. Denote the normalizing constant by

$$d_{in} = \int_{\mathbb{R}^d} \exp\left\{n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i)\right\}\pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}})\mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in})\mathrm{d}\mathbf{t}.$$

By definition,

$$\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A}) = \frac{1}{d_{in}}\exp\left\{n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i)\right\}\pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}})\mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in}).$$

It is sufficient to show that

$$\max_{i\in[n]}\int_{\mathbb{R}^d}(1 + \|\mathbf{t}\|_2^\alpha)\left|\exp\left\{n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i)\right\}\pi\left(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}\right)\mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in})\right.$$

$$\left. - e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}}\pi\left(\rho_n^{\frac{1}{2}}\mathbf{W}\mathbf{x}_{0i}\right)\right|\mathrm{d}\mathbf{t} = o_{\mathbb{P}_0}(1). \tag{27}$$

To see this, note that (15) in the manuscript can be rewritten as

$$\max_{i\in[n]}\frac{1}{d_{in}}\int(1 + \|\mathbf{t}\|_2^\alpha)\left|\exp\left\{n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i)\right\}\pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}})\mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in})\right.$$

$$\left. - \frac{d_{in}e^{-\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}/2}}{\det(2\pi\mathbf{G}_{0in}^{-1})^{1/2}}\right|\mathrm{d}\mathbf{t}$$

$$\leq \max_{i\in[n]} \frac{1}{d_{in}} \int \left(1+\|\mathbf{t}\|_2^\alpha\right) \left| \exp\left\{ n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i) \right\} \pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) \mathbb{1}(\mathbf{t}\in\widehat{\Theta}_{in}) \right.$$

$$\left. - e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}} \pi(\rho_n^{\frac{1}{2}}\mathbf{Wx}_{0i}) \right| \mathrm{d}\mathbf{t}$$

$$+ \max_{i\in[n]} \left| \frac{\pi(\mathbf{W}\rho_n^{1/2}\mathbf{x}_{0i})}{d_{in}} - \det(2\pi\mathbf{G}_{0in}^{-1})^{-1/2} \right| \int \left(1+\|\mathbf{t}\|_2^\alpha\right) e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}} \mathrm{d}\mathbf{t}.$$

Since (27) implies that $\max_{i\in[n]}|d_{in} - \det(2\pi\mathbf{G}_{0in}^{-1})^{1/2}\pi(\mathbf{W}\rho_n^{1/2}\mathbf{x}_{0i})| = o_{\mathbb{P}_0}(1)$ (by taking $\alpha=0$), it can be seen that (27) implies that the two terms on the right hand side of the previous display are $o_{\mathbb{P}_0}(1)$. Hence, we are left with establishing (27).

Let $\{\eta_n\}_{n=1}^\infty$ be a sequence to be determined later with $0 < \eta_n \to \infty$ and consider the following partition of $\mathbb{R}^d$:

$$\mathcal{A}_1 = \{\mathbf{t}\in\widehat{\Theta}_{in} : \|\mathbf{t}\|_2 \leq \eta_n\}, \qquad \mathcal{A}_2 = \{\mathbf{t}\in\widehat{\Theta}_{in} : \|\mathbf{t}\|_2 > \eta_n\}, \qquad \mathcal{A}_3 = \widehat{\Theta}_{in}^c.$$

We first consider the integral of (27) over $\mathcal{A}_3$. By definition of $\mathbb{1}(\mathbf{t}\in\widehat{\Theta}_{in})$, the integral over $\mathcal{A}_3$ can be bounded by

$$\max_{i\in[n]} \int_{\mathcal{A}_3} \left(1+\|\mathbf{t}\|_2^\alpha\right) e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}} \pi(\rho_n^{\frac{1}{2}}\mathbf{Wx}_{0i}) \mathrm{d}\mathbf{t}$$

$$\leq \int_{\mathcal{A}_3} \left(1+\|\mathbf{t}\|_2^\alpha\right) e^{-\min_{i\in[n]}\lambda_d(\mathbf{G}_{0in})\|\mathbf{t}\|_2^2/2} \pi(\rho_n^{1/2}\mathbf{Wx}_{0i}) \mathrm{d}\mathbf{t} \qquad (28)$$

$$\leq \int_{\mathcal{A}_3} \left(1+\|\mathbf{t}\|_2^\alpha\right) e^{-\lambda\|\mathbf{t}\|_2^2/2} \pi(\rho_n^{1/2}\mathbf{Wx}_{0i}) \mathrm{d}\mathbf{t} \to 0,$$

since $\mathcal{A}_3$ is shrinking to empty set and $\min_{i\in[n]}(\mathbf{G}_{0in}) \geq \lambda$ has been shown in the proof of Theorem 4 (see diaplay (16)). We next consider the integral of (27) over $\mathcal{A}_2$. Define the event

$$\mathcal{E}_{2n} = \left\{ \mathbf{A} : \max_{i\in[n]} \max_{\|\mathbf{x}_i\|_2\leq 1} \mathbf{s}^{\mathrm{T}} \frac{\partial^2\widetilde{M}_{in}}{\partial\mathbf{x}_i\partial\mathbf{x}_i^{\mathrm{T}}}(\mathbf{x}_i)\mathbf{s} \leq -\frac{\lambda}{2}\|\mathbf{s}\|_2^2 \quad \forall\mathbf{s}\in\mathbb{R}^d \right\}.$$

Note that by Lemma 14, Theorem 5.2 in Lei and Rinaldo (2015), and Weyl's inequality, with probability at least $1-n^{-c}$,

$$\min_{i\in[n]} \min_{\|\mathbf{x}_i\|_2\leq 1} \mathbf{s}^{\mathrm{T}} \left(-\frac{\partial^2\widetilde{M}_{in}}{\partial\mathbf{x}_i\partial\mathbf{x}_i^{\mathrm{T}}}(\mathbf{x}_i) \right) \mathbf{s} = \min_{i\in[n]} \min_{\|\mathbf{x}_i\|_2\leq 1} \mathbf{s}^{\mathrm{T}} \left(\frac{1}{n}\sum_{j=1}^n \left\{ \frac{1}{\widetilde{p}_{ij}} + \frac{1-A_{ij}}{(1-\mathbf{x}_i^{\mathrm{T}}\widetilde{\mathbf{x}}_j)^2} \right\} \widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}} \right) \mathbf{s}$$

$$\geq \frac{1}{\max_{i,j\in[n]}\widetilde{p}_{ij}} \frac{1}{n}\sum_{j=1}^n \mathbf{s}^{\mathrm{T}}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{s} \geq \frac{1}{n\rho_n}\mathbf{s}^{\mathrm{T}}\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{s} \geq \frac{1}{n\rho_n}\lambda_d(\mathbf{A})\|\mathbf{s}\|_2^2 \geq \frac{\lambda}{2}\|\mathbf{s}\|_2^2.$$

This shows that $\mathbb{P}_0(\mathcal{E}_{2n}) \geq 1-n^{-c}$ for all $n \geq N_{c,\delta,\lambda}$. By Taylor's expansion, for all $\mathbf{t}\in\widehat{\Theta}_{in}$, we have

$$n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i) = \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{W}\frac{\partial^2\widetilde{M}_{in}}{\partial\mathbf{x}_i\partial\mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i)\mathbf{Wt}, \qquad (29)$$

where $\bar{\mathbf{x}}_i = \widehat{\mathbf{x}}_i + \theta_i \mathbf{W}\mathbf{t}/\sqrt{n}$ for some $\theta_i \in [0,1]$ because the gradient of $\widetilde{M}_{in}$ evaluated at $\mathbf{x}_i = \widehat{\mathbf{x}}_i$ is zero by definition of the maximum surrogate likelihood estimator $\widehat{\mathbf{x}}_i$. Over this event, the integral of (27) over $\mathcal{A}_2$ can be upper bounded by

$$
\max_{i \in [n]} \int_{\mathcal{A}_2} (1 + \|\mathbf{t}\|_2^\alpha) \exp\left\{ \frac{1}{2} \mathbf{t}^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \frac{\partial^2 \widetilde{M}_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i) \mathbf{W}\mathbf{t} \right\} \pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) \mathrm{d}\mathbf{t}
$$

$$
+ \max_{i \in [n]} \int_{\mathcal{A}_2} (1 + \|\mathbf{t}\|_2^\alpha) e^{-\mathbf{t}^{\mathrm{T}} \mathbf{G}_{0in} \mathbf{t}/2} \pi(\rho_n^{1/2} \mathbf{W}\mathbf{x}_{0i}) \mathrm{d}\mathbf{t}
$$

$$
\leq C \int_{\mathcal{A}_2} (1 + \|\mathbf{t}\|_2^\alpha) \exp\left\{ \max_{i \in [n]} \max_{\|\mathbf{x}_i\|_2 \leq 1} \frac{1}{2} \mathbf{t}^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \frac{\partial^2 \widetilde{M}_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i) \mathbf{W}\mathbf{t} \right\} \mathrm{d}\mathbf{t}
$$

$$
+ \max_{i \in [n]} C \int_{\mathcal{A}_2} (1 + \|\mathbf{t}\|_2^\alpha) e^{-\mathbf{t}^{\mathrm{T}} \mathbf{G}_{0in} \mathbf{t}/2} \mathrm{d}\mathbf{t}
$$

$$
\leq 2C \int_{\|\mathbf{t}\|_2 > \eta_n} (1 + \|\mathbf{t}\|_2^\alpha) e^{-\lambda \|\mathbf{t}\|_2^2/4} \mathrm{d}\mathbf{t}.
$$

Denote the last line of the above display by ϵ_{2n}, then $\epsilon_{2n} \to 0$ because $\eta_n \to \infty$. It follows that

$$
\mathbb{P}_0 \left\{ \max_{i \in [n]} \int_{\mathcal{A}_2} (1 + \|\mathbf{t}\|_2^\alpha) \left| \exp\left\{ n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i) \right\} \pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) \mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in}) \right.
$$

$$
\left. - e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}} \mathbf{G}_{0in}\mathbf{t}} \pi(\rho_n^{\frac{1}{2}} \mathbf{W}\mathbf{x}_{0i}) \right| \mathrm{d}\mathbf{t} \geq \epsilon_{2n} \right\} \leq n^{-c}
$$

for all $n \geq N_{c,\delta,\lambda}$. Hence,

$$
\max_{i \in [n]} \int_{\mathcal{A}_2} (1 + \|\mathbf{t}\|_2^\alpha) \left| \exp\left\{ n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i) \right\} \pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{W}\mathbf{t}}{\sqrt{n}}) \mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in}) \right.
$$

$$
\left. - e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}} \mathbf{G}_{0in}\mathbf{t}} \pi(\rho_n^{\frac{1}{2}} \mathbf{W}\mathbf{x}_{0i}) \right| \mathrm{d}\mathbf{t} \xrightarrow{\mathbb{P}_0} 0. \tag{30}
$$

We next consider the integral of (27) over $\mathcal{A}_1$. Take $\eta_n = \min\{(n\rho_n/\log n)^{(1/8)}, \sqrt{(\log n)/\rho_n}\}$. Recall that $\mathbf{t} = \sqrt{n}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_i - \widehat{\mathbf{x}}_i)$, and $\max_{i \in [n]} \|\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}$ with probability at least $1 - n^{-c}$ by Theorem 4. Then

$$
\max_{i \in [n]} \|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 \leq \max_{i \in [n]} \|\mathbf{W}^{\mathrm{T}}\widehat{\mathbf{x}}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 + \max_{i \in [n]} \frac{\|\mathbf{t}\|_2}{\sqrt{n}} \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}
$$

with probability at least $1 - n^{-c}$ because $\eta_n/\sqrt{n} \leq \sqrt{(\log n)/(n\rho_n)}$, which also implies that there exists a constant $C_{c,\delta,\lambda} > 0$ (possibly depending on c, δ, λ), such that

$$
\{\mathbf{x}_i : \|\mathbf{t}\|_2 \leq \eta_n\} \subset \left\{ \mathbf{x}_i : \|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 \leq C_{c,\delta,\lambda}\sqrt{\frac{\log n}{n\rho_n}} \right\}
$$

with probability at least $1 - n^{-c}$. Define the event

$$\mathcal{E}_{1n} = \left\{ \mathbf{A} : \max_{i \in [n]} \sup_{\mathbf{x}_i : \|\mathbf{t}\|_2 \leq \eta_n} \left\| \mathbf{W}^{\mathrm{T}} \frac{\partial^2 \widetilde{M}_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\mathbf{x}_i)\mathbf{W} + \mathbf{G}_{0in} \right\|_2 \leq K_{c,\delta,\lambda}\sqrt{\frac{\log n}{n\rho_n}} \right\}$$

$$\cap \left\{ \mathbf{A} : \max_{i \in [n]} \|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 \leq K_{c,\delta,\lambda}\sqrt{\frac{\log n}{n\rho_n}} \right\}.$$

for an appropriate constant $K_{c,\delta,\lambda}$ depending on c, δ, λ. By Lemma 15, one can select $K_{c,\delta,\lambda}$ such that $\mathbb{P}_0(\mathcal{E}_{1n}) \geq 1 - n^{-c}$ for all $n \geq N_{c,\delta,\lambda}$. Then over the event $\mathcal{E}_{1n}$, by Taylor's expansion (29) and the mean-value theorem applied to the exponential function, we have

$$\max_{i \in [n]} \int_{\mathcal{A}_1} (1 + \|\mathbf{t}\|_2^\alpha) \left| \exp\left\{ n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i) \right\} \pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}})\mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in}) \right.$$

$$\left. - e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}}\pi(\rho_n^{\frac{1}{2}}\mathbf{Wx}_{0i}) \right| \mathrm{d}\mathbf{t}$$

$$= \max_{i \in [n]} \int_{\mathcal{A}_1} (1 + \|\mathbf{t}\|_2^\alpha) \left| \exp\left\{ \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}} \frac{\partial^2 \widetilde{M}_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i)\mathbf{Wt} \right\} \pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) \right.$$

$$\left. - e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}}\pi(\rho_n^{\frac{1}{2}}\mathbf{Wx}_{0i}) \right| \mathrm{d}\mathbf{t}$$

$$= \max_{i \in [n]} \int_{\mathcal{A}_1} (1 + \|\mathbf{t}\|_2^\alpha) \left| \exp\left\{ \frac{1}{2}\mathbf{t}^{\mathrm{T}} \left(\mathbf{W}^{\mathrm{T}} \frac{\partial^2 \widetilde{M}_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i)\mathbf{W} + \mathbf{G}_{0in} \right)\mathbf{t} \right\} - \frac{\pi(\rho_n^{1/2}\mathbf{Wx}_{0i})}{\pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}})} \right|$$

$$\times e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}}\pi\left(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}\right) \mathrm{d}\mathbf{t}$$

$$\leq \max_{i \in [n]} \int_{\mathcal{A}_1} (1 + \|\mathbf{t}\|_2^\alpha) \left\{ \left| \exp\left\{ \frac{1}{2}\mathbf{t}^{\mathrm{T}} \left(\mathbf{W}^{\mathrm{T}} \frac{\partial^2 \widetilde{M}_{in}}{\partial \mathbf{x}_i \partial \mathbf{x}_i^{\mathrm{T}}}(\bar{\mathbf{x}}_i)\mathbf{W} + \mathbf{G}_{0in} \right)\mathbf{t} \right\} - 1 \right| \right.$$

$$\left. + \left| 1 - \frac{\pi(\rho_n^{1/2}\mathbf{Wx}_{0i})}{\pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}})} \right| \right\} e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}}\pi\left(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}\right) \mathrm{d}\mathbf{t}$$

$$\leq \left(\exp\left\{ \frac{1}{2}K_{c,\delta,\lambda}\sqrt{\frac{\log n}{n\rho_n}}\eta_n^2 \right\} \frac{1}{2}K_{c,\delta,\lambda}\sqrt{\frac{\log n}{n\rho_n}}\eta_n^2 \right.$$

$$\left. + \max_{i \in [n]} \sup_{\mathbf{x}_i : \|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}} \left| 1 - \frac{\pi(\rho_n^{1/2}\mathbf{Wx}_{0i})}{\pi(\mathbf{x}_i)} \right| \right) \times C \int e^{-\lambda\|\mathbf{t}\|_2^2/2}\mathrm{d}\mathbf{t}.$$

Denote the last form of the above display by ϵ_{1n}. It is obvious that $\exp\left\{ \frac{1}{2}K_{c,\delta,\lambda}\sqrt{\frac{\log n}{n\rho_n}}\eta_n^2 \right\} \to 1$ (since $\eta_n = (n\rho_n/\log n)^{\frac{1}{8}}$). By the assumptions on $\pi(\mathbf{x}_i)$,

$$\max_{i \in [n]} \sup_{\mathbf{x}_i : \|\mathbf{W}^{\mathrm{T}}\mathbf{x}_i - \rho_n^{1/2}\mathbf{x}_{0i}\|_2 \lesssim_{c,\delta,\lambda} \sqrt{\frac{\log n}{n\rho_n}}} \left| 1 - \frac{\pi(\rho_n^{1/2}\mathbf{Wx}_{0i})}{\pi(\mathbf{x}_i)} \right| \to 0.$$

It follows that $\epsilon_{1n} \to 0$ as $n \to \infty$, and

$$\mathbb{P}_0 \left\{ \max_{i \in [n]} \int_{\mathcal{A}_1} (1 + \|\mathbf{t}\|_2^\alpha) \left| \exp\left\{ n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i) \right\} \pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) \mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in}) \right.\right.$$
$$\left.\left. - e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}} \pi(\rho_n^{\frac{1}{2}}\mathbf{Wx}_{0i}) \right| \mathrm{d}\mathbf{t} \geq \epsilon_{1n} \right\} \leq n^{-c},$$

for all $n \geq N_{c,\delta,\lambda}$. Hence,

$$\max_{i \in [n]} \int_{\mathcal{A}_1} (1 + \|\mathbf{t}\|_2^\alpha) \left| \exp\left\{ n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) - n\widetilde{M}_{in}(\widehat{\mathbf{x}}_i) \right\} \pi(\widehat{\mathbf{x}}_i + \frac{\mathbf{Wt}}{\sqrt{n}}) \mathbb{1}(\mathbf{t} \in \widehat{\Theta}_{in}) \right.$$
$$\left. - e^{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}} \pi(\rho_n^{\frac{1}{2}}\mathbf{Wx}_{0i}) \right| \mathrm{d}\mathbf{t} \xrightarrow{\mathbb{P}_0} 0. \tag{31}$$

The proof of (27) is completed by combining (28), (30), and (31). ∎

B.5 Proof of Corollary 12

Proof We first show the convergence of the mean and covariance of $\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A})$, which is a direct consequence of Theorem 11:

$$\max_{i \in [n]} \left\| \int \mathbf{t}\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A})\mathrm{d}\mathbf{t} \right\|_2 = \max_{i \in [n]} \left\| \int \mathbf{t}\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A})\mathrm{d}\mathbf{t} - \int \mathbf{t}\frac{e^{-\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}/2}}{\det(2\pi\mathbf{G}_{0in}^{-1})^{1/2}}\mathrm{d}\mathbf{t} \right\|_2$$
$$\leq \max_{i \in [n]} \int \|\mathbf{t}\|_2 \left| \widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A}) - \frac{e^{-\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}/2}}{\det(2\pi\mathbf{G}_{0in}^{-1})^{1/2}} \right| \mathrm{d}\mathbf{t} \xrightarrow{\mathbb{P}_0} 0,$$

and

$$\max_{i \in [n]} \left\| \int \mathbf{t}\mathbf{t}^{\mathrm{T}}\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A})\mathrm{d}\mathbf{t} - \mathbf{G}_{0in}^{-1} \right\|_2 = \max_{i \in [n]} \left\| \int \mathbf{t}\mathbf{t}^{\mathrm{T}}\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A})\mathrm{d}\mathbf{t} - \int \mathbf{t}\mathbf{t}^{\mathrm{T}}\frac{e^{-\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}/2}}{\det(2\pi\mathbf{G}_{0in}^{-1})^{1/2}}\mathrm{d}\mathbf{t} \right\|_2$$
$$\leq \max_{i \in [n]} \int \|\mathbf{t}\|_2^2 \left| \widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A}) - \frac{e^{-\mathbf{t}^{\mathrm{T}}\mathbf{G}_{0in}\mathbf{t}/2}}{\det(2\pi\mathbf{G}_{0in}^{-1})^{1/2}} \right| \mathrm{d}\mathbf{t} \xrightarrow{\mathbb{P}_0} 0.$$

Now

$$\max_{i \in [n]} \|\sqrt{n}(\mathbf{x}_i^* - \widehat{\mathbf{x}}_i)\|_2 = \max_{i \in [n]} \left\| \int \sqrt{n}(\mathbf{x}_i - \widehat{\mathbf{x}}_i)\widetilde{\pi}_{in}(\mathbf{x}_i \mid \mathbf{A})\mathrm{d}\mathbf{x}_i \right\|_2$$
$$= \max_{i \in [n]} \left\| \int \mathbf{t}\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A})\mathrm{d}\mathbf{t} \right\|_2$$
$$= o_{\mathbb{P}_0}(1),$$

then by Theorem 4 and Slutsky's Theorem, $\sqrt{n}\mathbf{G}_{0in}^{1/2}(\mathbf{W}^{\mathrm{T}}\mathbf{x}_i^* - \rho_n^{1/2}\mathbf{x}_{0i}) \xrightarrow{\mathcal{L}} \mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d)$. Also,

$$\max_{i \in [n]} \left\| n\mathbf{W}^{\mathrm{T}}\mathbf{\Sigma}_{in}^*\mathbf{W} - \mathbf{G}_{0in}^{-1} \right\|_2$$

$$\begin{aligned}
&= \max_{i \in [n]} \left\| \int n\mathbf{W}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_i^*)(\mathbf{x}_i - \mathbf{x}_i^*)^{\mathrm{T}}\mathbf{W}\widetilde{\pi}_{in}(\mathbf{x}_i \mid \mathbf{A})\mathrm{d}\mathbf{x}_i - \mathbf{G}_{0in}^{-1} \right\|_2 \\
&= \max_{i \in [n]} \left\| \int n\mathbf{W}^{\mathrm{T}}(\mathbf{x}_i - \widehat{\mathbf{x}}_i + \widehat{\mathbf{x}}_i - \mathbf{x}_i^*)(\mathbf{x}_i - \widehat{\mathbf{x}}_i + \widehat{\mathbf{x}}_i - \mathbf{x}_i^*)^{\mathrm{T}}\mathbf{W}\widetilde{\pi}_{in}(\mathbf{x}_i \mid \mathbf{A})\mathrm{d}\mathbf{x}_i - \mathbf{G}_{0in}^{-1} \right\|_2 \\
&\leq \max_{i \in [n]} \left\| \int \mathbf{t}\mathbf{t}^{\mathrm{T}}\widetilde{\pi}_{in}^*(\mathbf{t} \mid \mathbf{A})\mathrm{d}\mathbf{t} - \mathbf{G}_{0in}^{-1} \right\|_2 + o_{\mathbb{P}_0}(1) \\
&= o_{\mathbb{P}_0}(1).
\end{aligned}$$

Note that $\mathbf{G}_{0in}$ is finite and positive definite. By continuous mapping theorem,

$$(\rho_n^{1/2}\mathbf{W}\mathbf{x}_i - \mathbf{x}_i^*)^{\mathrm{T}}(\mathbf{\Sigma}_{in}^*)^{-1}(\rho_n^{1/2}\mathbf{W}\mathbf{x}_i - \mathbf{x}_i^*) \xrightarrow{\mathcal{L}} \chi_d^2,$$

so $\mathbb{P}_0\{(\rho_n^{1/2}\mathbf{W}\mathbf{x}_i - \mathbf{x}_i^*)^{\mathrm{T}}(\mathbf{\Sigma}_{in}^*)^{-1}(\rho_n^{1/2}\mathbf{W}\mathbf{x}_i - \mathbf{x}_i^*) \leq q_{1-\alpha}\} \to 1 - \alpha$.

We now focus on the last assertion. By the previous proof, we know that $\max_{i \in [n]} \|\mathbf{x}^* - \widehat{\mathbf{x}}_i\|_2^2 = o_{\mathbb{P}_0}(1/n)$. It follows directly that

$$\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \|\mathbf{x}^* - \widehat{\mathbf{x}}_i\|_2^2 \leq n \max_{i \in [n]} \|\mathbf{x}^* - \widehat{\mathbf{x}}_i\|_2^2 = o_{\mathbb{P}_0}(1).$$

Therefore, by Theorem 4 and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\|\mathbf{X}^*\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 &= \|\widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}^2 + \|\mathbf{X}^*\mathbf{W} - \widehat{\mathbf{X}}\mathbf{W}\|_{\mathrm{F}}^2 \\
&\quad + 2\left\langle \widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0, \mathbf{X}^*\mathbf{W} - \widehat{\mathbf{X}}\mathbf{W} \right\rangle_{\mathrm{F}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}(\mathbf{G}_{0in}^{-1}) + o_{\mathbb{P}_0}(1) + O\left(\|\widehat{\mathbf{X}}\mathbf{W} - \rho_n^{1/2}\mathbf{X}_0\|_{\mathrm{F}}\|\mathbf{X}^*\mathbf{W} - \widehat{\mathbf{X}}\mathbf{W}\|_{\mathrm{F}} \right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}(\mathbf{G}_{0in}^{-1}) + o_{\mathbb{P}_0}(1),
\end{aligned}$$

where $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ denotes the Frobenius inner product between matrices. The proof is thus completed. ∎

Appendix C. Proof of the Convergence of the Stochastic Gradient Descent

Lemma 19 (Lemma A.5 in Mairal, 2013) *Let $(a_t)_{t \geq 1}, (b_t)_{t \geq 1}$ be two non-negative real sequences. Assume that $\sum_{t=1}^{\infty} a_t b_t$ converges and $\sum_{t=1}^{\infty} a_t$ diverges, and $|b_{t+1} - b_t| \leq Ka_t$ for some constant $K \geq 0$. Then b_t converges to 0.*

Lemma 20 (Lemma 2 in Li and Orabona, 2019) *Let $a_0 > 0$, $a_i \geq 0$, $i = 1, \dots, T$ and $\beta > 1$. Then $\sum_{t=1}^{T} \frac{a_t}{(a_0 + \sum_{i=1}^{t} a_i)^{\beta}} < \frac{1}{(\beta-1)a_0^{\beta-1}}$.*

Lemma 21 (Lemma 3 in Li and Orabona, 2019) *Let $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable whose minimum is attained at $\mathbf{x} = \mathbf{x}^*$ and suppose there exists a constant $L > 0$, such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,*

$$\left\| \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}) - \frac{\partial f}{\partial \mathbf{x}}(\mathbf{y}) \right\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2.$$

Suppose $\mathbf{g}(\mathbf{x}, \mathbf{z})$ is a function of a random vector $\mathbf{z}$, such that $\mathbb{E}_z \mathbf{g}(\mathbf{x}, \mathbf{z}) = \partial f(\mathbf{x})/\partial \mathbf{x}$. Let $(\mathbf{z}_t)_{t \geq 1}$ be a sequence of independent and identically distributed (i.i.d.) copies of $\mathbf{z}$. Consider a sequence of iterates $\mathbf{x}^{(t)}$ generated by

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \mathbf{H}_t \mathbf{g}(\mathbf{x}^{(t)}, \mathbf{z}_t),$$

where $\mathbf{H}_t \in \mathbb{R}^{d \times d}$ is a step-size matrix for the tth iteration. Then the sequence $(\mathbf{x}^{(t)})_{t \geq 1}$ satisfies the following inequality:

$$\mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_N} \left[\sum_{t=1}^{N} \left\langle \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}^{(t)}), \mathbf{H}_t \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}^{(t)}) \right\rangle \right]$$

$$\leq f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*) + \frac{L}{2} \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_N} \left\{ \sum_{t=1}^{N} \| \mathbf{H}_t \mathbf{g}(\mathbf{x}^{(t)}, \mathbf{z}_t) \|^2 \right\}.$$

Proof of Theorem 9. The proof is similar to Theorem 1 in Li and Orabona (2019), with some slight modifications. In the setting here, the expectation is taken with respect to the randomness of the stochastic gradient descent conditioned on the adjacency matrix, that is, the data and the ASE are viewed as deterministic. Here, we suppress the subscript $i \in [n]$ and use $\mathbf{x}^{(t)}$ to denote the tth iterate in the optimization, and $\widehat{\mathbf{x}}$ the maximizer of the average surrogate log-likelihood function $\widetilde{M}_{in}(\mathbf{x}) := (1/n)\widetilde{\ell}_{in}(\mathbf{x})$.

For the surrogate log-likelihood function, by the computation of the gradient and Hessian of $\widetilde{M}_{in}$ in the proof of Theorem 4, they are bounded over $\{\mathbf{x}_i : \|\mathbf{x}_i\|_2 \leq 1\}$ when $\max_j \|\widetilde{\mathbf{x}}_j\|_2 < 1$. So both $\widetilde{M}_{in}(\mathbf{x})$ and its gradient are Lipschitz in $\{\mathbf{x} \in \mathrm{R}^d : \|\mathbf{x}\|_2 \leq 1\}$ by the mean value theorem. Let C_1 and C_2 be the Lipschitz constants for $\widetilde{M}_{in}(\mathbf{x})$ and its gradient, respectively. In the context of Section 3.2, the random vector $\mathbf{z}_t$ corresponds to the randomly generated indices $(j_1^{(t)}, \ldots, j_s^{(t)})$ in a single iteration of the mini-batch SGD algorithm, and $\mathbf{g}(\mathbf{x}^{(t)}, \mathbf{z}_t)$ takes the form

$$\mathbf{g}(\mathbf{x}^{(t)}, \mathbf{z}_t) = \frac{1}{s} \sum_{k=1}^{s} \frac{\partial m_i}{\partial \mathbf{x}}(\mathbf{x}^{(t)}, j_k^{(t)}).$$

It is clear that $\mathbb{E}_{\mathbf{z}_t} \mathbf{g}(\mathbf{x}^{(t)}, \mathbf{z}_t)$ coincides with the gradient of $\widetilde{M}_{in}(\mathbf{x}^{(t)})$. Also, for the stochastic gradient $\mathbf{g}(\mathbf{x}^{(t)}, z_t)$, it is easy to see that $\|\mathbf{g}(\mathbf{x}^{(t)}, \mathbf{z}_t) - \nabla \widetilde{M}_{in}(\mathbf{x}^{(t)})\|_2 \leq C_3$ for all $\mathbf{x}^{(t)} \in B(\mathbf{0}_d, 1)$.

Observe that

$$\sum_{t=1}^{\infty} \|\alpha_t \mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 = \sum_{t=1}^{\infty} \alpha_{t+1}^2 \|\mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 + \sum_{t=1}^{\infty} (\alpha_t^2 - \alpha_{t+1}^2) \|\mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2$$

$$\leq \frac{a_0^2}{2\epsilon b_0^{2\epsilon}} + \max_{t \geq 1} \|\mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 \sum_{t=1}^{\infty} (\alpha_t^2 - \alpha_{t+1}^2)$$

$$\leq \frac{a_0^2}{2\epsilon b_0^{2\epsilon}} + \max_{t \geq 1} \|\mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 \alpha_1^2$$

$$\leq \frac{a_0^2}{2\epsilon b_0^{2\epsilon}} + 2\alpha_1^2 \max_{t \geq 1} \left(\|\|\nabla \widetilde{M}_{in}(\mathbf{x}^{(t)})\|_2^2 + \|\nabla \widetilde{M}_{in}(\mathbf{x}^{(t)}) - \mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 \right)$$

$$\leq \frac{a_0^2}{2\epsilon b_0^{2\epsilon}} + \frac{2a_0^2}{b_0^{1+2\epsilon}}(C_1^2 + C_3^2) < \infty,$$

where in the first inequality we have used Lemma 20, in the third one the elementary inequality $\|\mathbf{x} + \mathbf{y}\|_2^2 \leq 2\|\mathbf{x}\|_2^2 + 2\|\mathbf{y}\|_2^2$. Therefore, for all $m \in \mathbb{N}_+$, by Cauchy–Schwarz inequality, we have

$$\left\| \mathbf{x}^{(N+m)} - \mathbf{x}^{(N)} \right\|_2^2 = \left\| \sum_{t=N}^{N+m-1} \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \right\|_2^2 \leq \sum_{t=N}^{N+m-1} \left\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \right\|_2^2$$

$$\leq \sum_{t=N}^{N+m-1} \left\| \alpha_t \mathbf{g}(\mathbf{x}^{(t)}, z_t) \right\|_2^2,$$

and the previous infinite sum being finite implies that $\lim_{N \to \infty} \left\| \mathbf{x}^{(N+m)} - \mathbf{x}^{(N)} \right\|_2 = 0$ a.s., that is, $\{\mathbf{x}^{(t)}\}_t$ forms a Cauchy sequence, and thus converges to some point $\mathbf{x}^* \in B(0,1)$ a.s.. Note that $\mathbf{x}^*$ is a random variable with respect to the randomness of z_t. Next we need to show that $\mathbf{x}^*$ is indeed the maximizer of the surrogate log-likelihood function.

By Lemma 21, taking the limit $T \to \infty$ and exchanging the expectation and the limits due to non-negative terms, we have

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \alpha_t \left\| \nabla \widetilde{M}_{in}(\mathbf{x}^{(t)}) \right\|_2^2 \right] \leq \widetilde{M}_{in}(\mathbf{x}^*) - \widetilde{M}_{in}(\mathbf{x}_1) + \frac{C_2}{2} \mathbb{E}\left[\sum_{t=1}^{\infty} \|\alpha_t \mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 \right].$$

With the right hand side being finite, we have

$$\sum_{t=1}^{\infty} \alpha_t \left\| \nabla \widetilde{M}_{in}(\mathbf{x}^{(t)}) \right\|_2^2 < \infty.$$

Observe that by definition,

$$\sup_{z_t, \mathbf{x}^{(t)}} \left\| \alpha_t \mathbf{g}(\mathbf{x}^{(t)}, z_t) \right\|_2 \leq \frac{a_0}{(b_0)^{1/2+\epsilon}} \sup_{z_t, \mathbf{x}^{(t)}} \left\| \mathbf{g}(\mathbf{x}^{(t)}, z_t) \right\|_2 < \infty,$$

that is, the updating of the iterate is bounded. By assumption, the MSLE $\widehat{\mathbf{x}}$ is in the interior of the feasible region. So there exists an integer m^* such that for all $t \in \mathbb{N}_+$, the number of times that step-halving in the algorithm is called is no greater than m^*. This implies that

$$\frac{1}{m^*} a_0 \left[b_0 + \sum_{i=1}^{t-1} \|\mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 \right]^{-(1/2+\epsilon)} \leq \alpha_t \leq a_0 \left[b_0 + \sum_{i=1}^{t-1} \|\mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 \right]^{-(1/2+\epsilon)}$$

for all $t \in \mathbb{N}_+$, which further implies that

$$\sum_{t=1}^{\infty} \alpha_t \geq \frac{1}{m^*} \sum_{t=1}^{\infty} a_0 \left[b_0 + \sum_{i=1}^{t-1} \|\mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2^2 \right]^{-(1/2+\epsilon)}$$

$$\geq \frac{1}{m^*} \sum_{t=1}^{\infty} a_0 \left[b_0 + 2(t-1)(C_1^2 + C_3^2) \right]^{-(1/2+\epsilon)} = \infty.$$

Using the fact that both $\widetilde{M}_{in}(\mathbf{x})$ and $\nabla \widetilde{M}_{in}(\mathbf{x})$ are Lipschitz, we also have

$$\left| \|\nabla \widetilde{M}_{in}(\mathbf{x}_{t+1})\|_2^2 - \|\nabla \widetilde{M}_{in}(\mathbf{x}^{(t)})\|_2^2 \right|$$

$$= \left(\|\nabla \widetilde{M}_{in}(\mathbf{x}_{t+1})\|_2 + \|\nabla \widetilde{M}_{in}(\mathbf{x}^{(t)})\|_2 \right) \cdot \left| \|\nabla \widetilde{M}_{in}(\mathbf{x}_{t+1})\|_2 - \|\nabla \widetilde{M}_{in}(\mathbf{x}^{(t)})\|_2 \right|$$

$$\leq 2C_1 C_2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2 = 2C_1 C_2 \|\alpha_t \mathbf{g}(\mathbf{x}^{(t)}, z_t)\|_2 \leq 2C_1 C_2 (C_1 + C_3)\alpha_t.$$

Hence, we can use Lemma 19 to obtain that $\lim_{t \to \infty} \|\nabla \widetilde{M}_{in}(\mathbf{x}^{(t)})\|_2 = 0$ a.s.. The continuity of $\nabla M_{in}(\mathbf{x})$ implies that $\mathbf{x}^{(t)} \to \hat{\mathbf{x}}$ a.s.. ∎

Appendix D. Additional Implementation Details

D.1 Additional Details of the Algorithms

This subsection provides the detailed Metropolis–Hastings sampler for computing the joint posterior distribution $\pi_n(\mathbf{X} \mid \mathbf{A})$ using the surrogate likelihood function. For each $i \in [n]$, we use the normal random walk truncated in the unit ball as the proposal distribution, with the covariance matrix being the inverse of

$$n\widetilde{\mathbf{G}}_{in} = \sum_{j=1}^{n} \frac{\widetilde{\mathbf{x}}_j \widetilde{\mathbf{x}}_j^{\mathrm{T}}}{\widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j (1 - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j)}.$$

The above covariance matrix is the plug-in estimator of the asymptotic covariance matrix of the Bernstein–von Mises limit distribution. Below, we provide the detailed Metropolis–Hastings sampler in the algorithm below. The computation of the posterior distribution of the entire latent position matrix $\mathbf{X}$ can be done by a parallelization over $i \in [n]$.

D.2 Convergence Diagnostics of the Metropolis–Hastings Sampler

In this subsection, we provide some convergence diagnostics of Metropolis–Hastings sampler. Specifically, we choose one realization of the simulated data in the case of the stochastic block model with $d = 2$ and $n = 2000$ (Section 5.3 of the manuscript). The parameters of this random dot product graph are the entries of a 2000×2 matrix, so we get $2000 \times 2 = 4000$ Markov chains as the output of Metropolis–Hastings sampler. The total number of iterations in one Markov chain is 2000, where we discard the first 2000 as burn-in and apply a thinning of 5 to the rest, resulting in a chain of length 200. To diagnose convergence, we use

---

Algorithm 2 Metropolis–Hastings sampler for computing the posterior distribution of $\mathbf{X}$.

1: **Input:** The adjacency matrix $\mathbf{A} = [A_{ij}]_{n \times n}$;
2:     The embedding dimension d;
3:     The tuning parameter σ;
4:     Number of burn-in iterations B;
5:     Number of post-burn-in samples n_{mc};
6:     Thinning size b.
7: Compute the spectral decomposition of the adjacency matrix

$$\mathbf{A} = \sum_{i=1}^{n} \widehat{\lambda}_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_j^{\mathrm{T}},$$

   where $|\widehat{\lambda}_1| \geq |\widehat{\lambda}_2| \geq \ldots \geq |\widehat{\lambda}_n|$, and $\widehat{\mathbf{u}}_i^{\mathrm{T}} \widehat{\mathbf{u}}_j = \mathbb{1}(i = j)$ for all $i, j \in [n]$.
8: Compute the adjacency spectral embedding:

$$\widetilde{\mathbf{X}} = \widehat{\mathbf{X}}^{\mathrm{ASE}} = [\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_d] \cdot \mathrm{diag}(|\widehat{\lambda}_1|^{1/2}, \ldots, |\widehat{\lambda}_d|^{1/2}),$$

   and write $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d}$. Let $= \widetilde{p}_{ij} = \widetilde{\mathbf{x}}_i^{\mathrm{T}} \widetilde{\mathbf{x}}_j$ for all $i, j \in [n]$.
9: For $i = 1, 2, \ldots, n$
10:     Initialize $\mathbf{x}_i^{(1)} = \widetilde{\mathbf{x}}_i$.
11:     For $t = 2$ to $B + n_{\mathrm{mc}} \times b$
12:         Generate $\mathbf{x}_i' \sim N\left(\mathbf{x}_i^{(t)}, \sigma^2 \widetilde{\mathbf{G}}_{in}^{-1}/n\right) \cdot \mathbb{1}(\|\mathbf{x}_i\|_2 < 1)$.
13:         Generate $\alpha_t \sim \mathrm{Unif}(0, 1)$.
14:         If $\log \alpha_t < \widetilde{\ell}_{in}(\mathbf{x}_i') - \widetilde{\ell}_{in}(\mathbf{x}_i^{(t)}) + \log \pi(\mathbf{x}_i') - \log \pi(\mathbf{x}_i^{(t)})$
15:             Set $\mathbf{x}_i^{(t+1)} \leftarrow \mathbf{x}_i'$;
16:         Else
17:             Set $\mathbf{x}_i^{(t+1)} \leftarrow \mathbf{x}_i^{(t)}$.
18:         End If
19:     End For
20: End For
21: **Output:** $\mathbf{X}^{(B+1+b \times N)}$ for $N = 1, 2, \ldots, \lceil (n_{\mathrm{mc}} - 1)/b \rceil$, where $\mathbf{X}^{(t)} = [\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}]^{\mathrm{T}}$.

`coda::heidel.diag()` in R, which uses the Cramer–von Mises statistic to test the null hypothesis that the sampled values come from a stationary distribution.

Below, Fig. 5 presents the numerical diagnostics results. From the histogram of the 4000 p-values from the output of `coda::heidel.diag()` applied to the 4000 Markov chains, we see that there are very few p-values that are less than 0.05 (only 36 among the 4000 p-values in this trial). Furthermore, with different trials of Metropolis–Hastings sampler, the specific parameters which give the small p-values are different. So we can say that the occurrence of some small p-values is very likely due to the randomness in the data and in the Metropolis–Hastings sampler. A histogram of the accept rates from the Metropolis–Hastings algorithm of the 2000 vertices is provided as well. To investigate more closely, the trace plot and auto-correlation function (ACF) plot of the second coordinate of the 808th vertex which

gives a p-value smaller than 0.05 in this trial are provided. We can see that although it gives a small p-value, the trace plot and the ACF plot of the Metropolis–Hastings sample are not too abnormal.
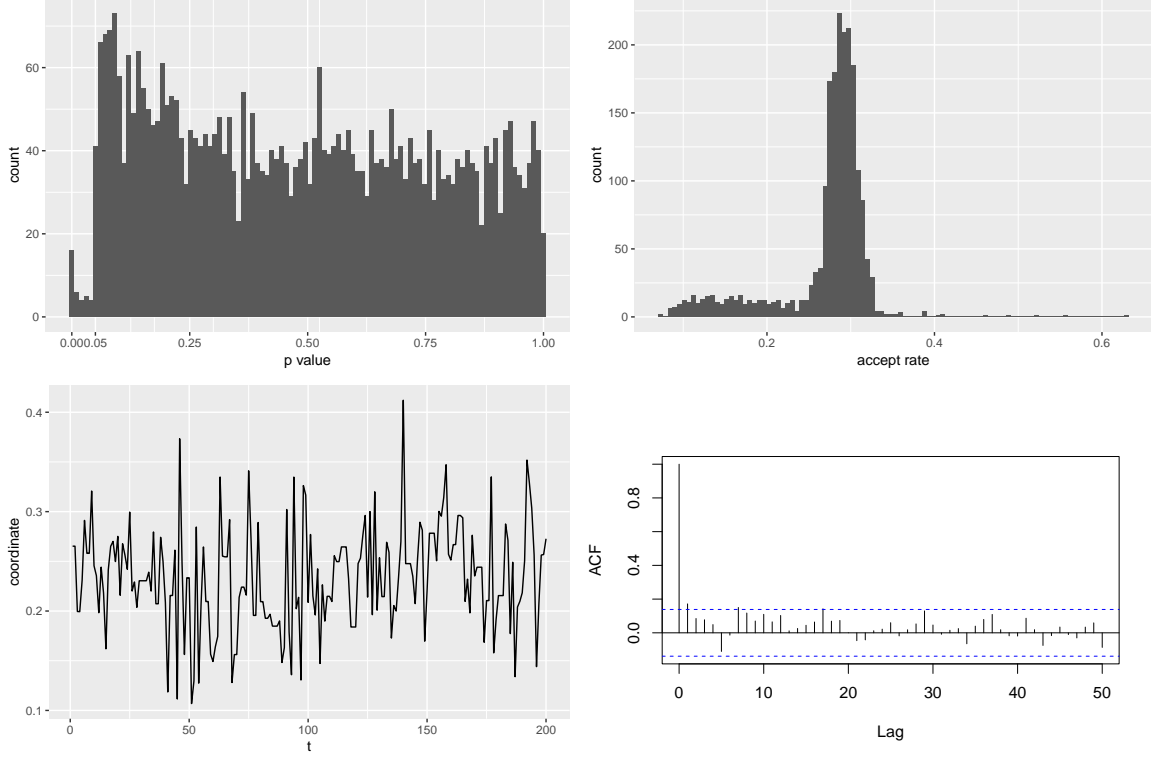

Figure 5: Convergence diagnostics for the simulation example in Section 5.3 of the manuscript. Top left panel: histogram of 4000 p-values. Top right panel: histogram of 2000 accept rates. Bottom left panel: Trace plot of a parameter whose Metropolis–Hastings sample gives a p-value less than 0.05. Bottom right panel: ACF plot of a parameter whose Metropolis–Hastings sample gives a p-value less than 0.05.

Next, we invectigate the convergence of the Metropolis–Hastings sampler in the Wikipedia graph data set (Section 5.4 of the manuscript). For each d, there are $1382 \times d$ parameters to estimate, so we get $1382 \times d$ markov chains as the output of Metropolis–Hastings sampler. The total number of iterations in one Metropolis–Hastings sampler is $4000(2d + 1)$, where we discard the first half as burn-in and apply a thinning of $4d$, resulting in a chain of length slightly more than 1000.

For $d = 1, \ldots, 15$, the histograms of 1382 accept rates and of $1382 \times d$ p-values are provided in the upper and lower panel of Fig. 7, respectively.

To investigate more closely, the trace plots and autocorrelation function (ACF) plots of two chains which give p-values smaller than 0.05 are provided, as in Fig. 8.
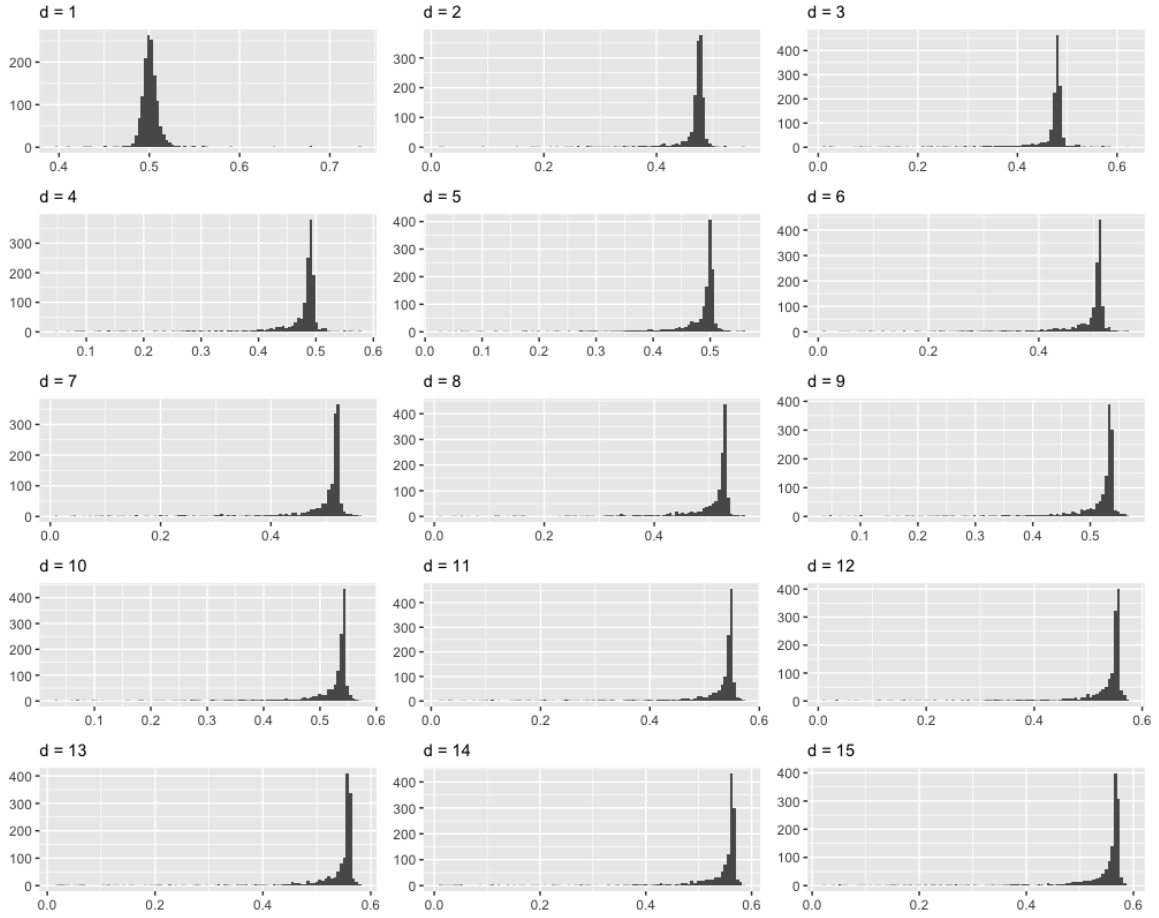
56

Figure 6: Convergence diagnostics for the Wikipedia graph data example in Section 5.4 of the manuscript: Histograms of accept rates, where the horizontal axis represents accept rates and the vertical axis represents counts.
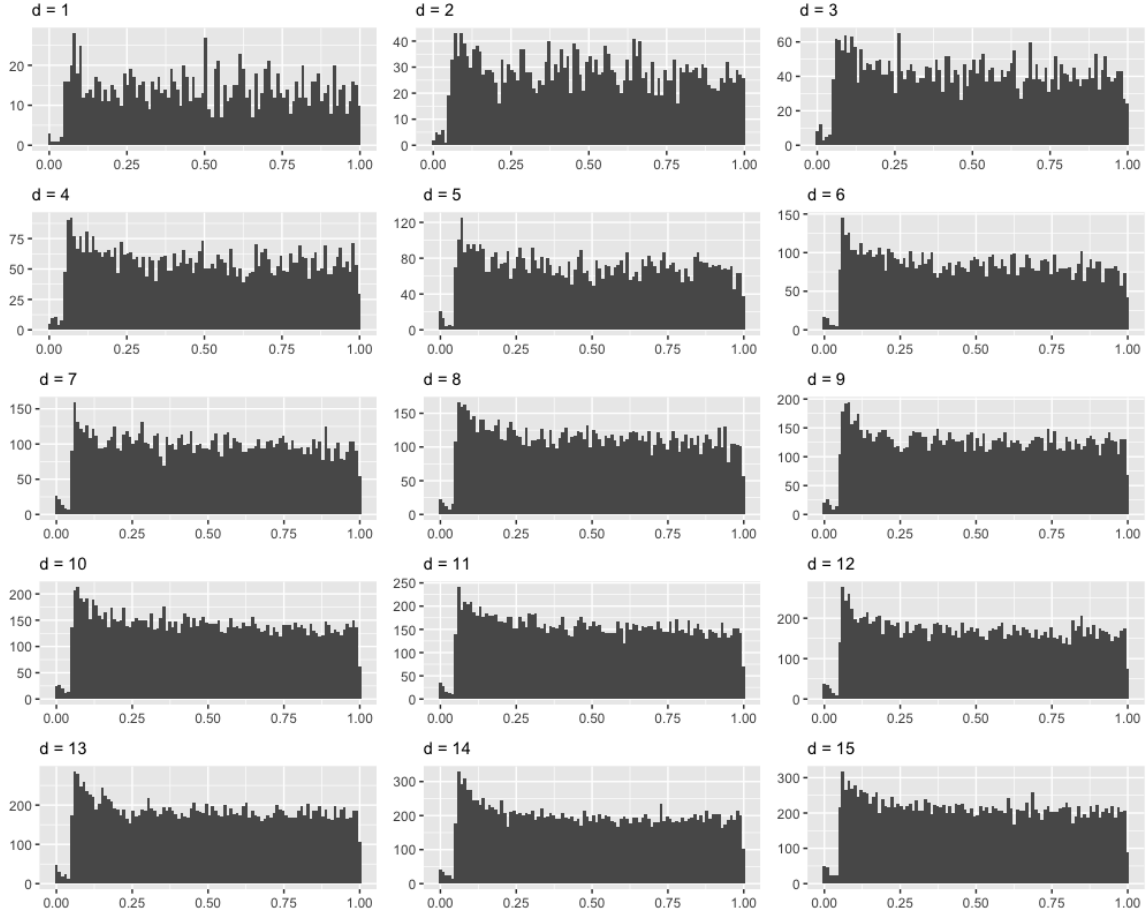
Figure 7: Convergence diagnostics for the Wikipedia graph data example in Section 5.4 of the manuscript. Top panel: histograms of accept rates, where the horizontal axis represents accept rates and the vertical axis represents counts. Bottom panel: Histograms of p-values, where the horizontal axis represents p-values and the vertical axis represents counts.
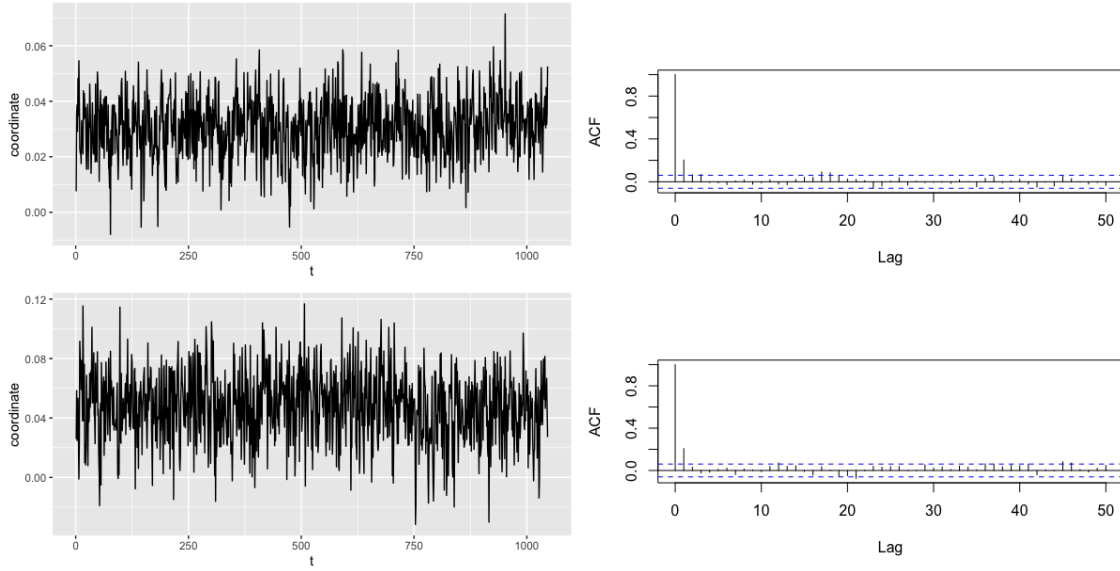
Figure 8: Convergence diagnostics for the Wikipedia graph data example in Section 5.4 of the manuscript. Top left panel: Trace plot of the Markov chain of the first coordinate of the $354th$ vertex with p-value $= 0.0019$, $d = 11$. Top right panel: ACF plot of the Markov chain of the first coordinate of the $354th$ vertex, $d = 11$. Bottom left panel: Trace plot of the Markov chain of the tenth coordinate of the $14th$ vertex with p-value $= 0.0004$, $d = 11$. Bottom right panel: ACF plot of the Markov chain of the tenth coordinate of the $14th$ vertex, $d = 11$.

References

Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016. doi: 10.1109/TIT.2015.2490670.

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, 48(3):1452 – 1474, 2020. doi: 10.1214/19-AOS1854. URL `https://doi.org/10.1214/19-AOS1854`.

Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 36—-43, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595932151. doi: 10.1145/1134271.1134277. URL `https://doi.org/10.1145/1134271.1134277`.

Edo M Airoldi, David Blei, Stephen Fienberg, and Eric Xing. Mixed membership stochastic blockmodels. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL `https://proceedings.neurips.cc/paper/2008/file/8613985ec49eb8f757ae6439e879bb2a-Paper.pdf`.

Arash A. Amini, Aiyou Chen, Peter J. Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013. doi: 10.1214/13-AOS1138. URL `https://doi.org/10.1214/13-AOS1138`.

Avanti Athreya, Carey E Priebe, Minh Tang, Vince Lyzinski, David J Marchette, and Daniel L Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1):1–18, 2016.

Avanti Athreya, Donniell E. Fishkind, Keith Levin, Vince Lyzinski, Youngser Park, Yichen Qin, Daniel L. Sussman, Minh Tang, Joshua T. Vogelstein, and Carey E. Priebe. Statistical inference on random dot product graphs: A survey. *Journal of Machine Learning Research*, 18, 09 2017.

Avanti Athreya, Joshua Cape, and Minh Tang. Eigenvalues of stochastic blockmodel graphs and random graphs with low-rank edge probability matrices. *Sankhya A: The Indian Journal of Statistics*, 84(1):36–63, 2022.

Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics*. Pearson Prentice Hall, 2007.

Peter J. Bickel and Purnamrita Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(1):253–273, 05 2015. ISSN 1369-7412. doi: 10.1111/rssb.12117. URL `https://doi.org/10.1111/rssb.12117`.

N. Binkiewicz, J. T. Vogelstein, and K. Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 03 2017. ISSN 0006-3444. doi: 10.1093/biomet/asx008. URL https://doi.org/10.1093/biomet/asx008.

Victor Chernozhukov and Han Hong. An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003. ISSN 0304-4076. doi: https://doi.org/10.1016/S0304-4076(03)00100-3. URL https://www.sciencedirect.com/science/article/pii/S0304407603001003.

Victor H. de la Pena and S. J. Montgomery-Smith. Decoupling Inequalities for the Tail Probabilities of Multivariate *U*-Statistics. *The Annals of Probability*, 23(2):806 – 816, 1995. doi: 10.1214/aop/1176988291. URL https://doi.org/10.1214/aop/1176988291.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936. doi: 10.1007/BF02288367.

Paul Erdős, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624 – 2652, 2015. doi: 10.1214/15-AOS1354. URL https://doi.org/10.1214/15-AOS1354.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002. doi: 10.1198/016214502388618906. URL https://doi.org/10.1198/016214502388618906.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. ISSN 0378-8733. doi: https://doi.org/10.1016/0378-8733(83)90021-7. URL https://www.sciencedirect.com/science/article/pii/0378873383900217.

Svante Janson and Persi Diaconis. Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni. Serie VII*, 28:33–61, 2008.

Jiashun Jin. Fast community detection by SCORE. *The Annals of Statistics*, 43(1):57–89, 2015. doi: 10.1214/14-AOS1265. URL https://doi.org/10.1214/14-AOS1265.

Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83:016107, Jan 2011. doi: 10.1103/PhysRevE.83.016107. URL https://link.aps.org/doi/10.1103/PhysRevE.83.016107.

B.J.K. Kleijn and A.W. van der Vaart. The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6(none):354 – 381, 2012. doi: 10.1214/12-EJS675. URL https://doi.org/10.1214/12-EJS675.

Michael R Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer New York, New York, NY, 2008. ISBN 978-0-387-74978-5. doi: 10.1007/978-0-387-74978-5.

Nicola Lacetera, Mario Macis, and Angelo Mele. Viral altruism? charitable giving and social contagion in online networks. *Sociological Science*, 3(11):202–238, 2016. ISSN 2330-6696. doi: 10.15195/v3.a11. URL http://dx.doi.org/10.15195/v3.a11.

Can M. Le, Elizaveta Levina, and Roman Vershynin. Optimization via low-rank approximation for community detection in networks. *The Annals of Statistics*, 44(1):373–400, 2016. doi: 10.1214/15-AOS1360. URL https://doi.org/10.1214/15-AOS1360.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215 – 237, 2015. doi: 10.1214/14-AOS1274. URL https://doi.org/10.1214/14-AOS1274.

Jinming Li, Gongjun Xu, and Ji Zhu. Statistical inference on latent space models for network data. *arXiv preprint arXiv:2312.06605*, 2023.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 983–992. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/li19c.html.

Vince Lyzinski, Daniel L. Sussman, Minh Tang, Avanti Athreya, and Carey E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905 – 2922, 2014. doi: 10.1214/14-EJS978. URL https://doi.org/10.1214/14-EJS978.

Vince Lyzinski, Minh Tang, Avanti Athreya, Youngser Park, and Carey E. Priebe. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering*, 4(1):13–26, 2017. doi: 10.1109/TNSE.2016.2634322.

Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/4da04049a062f5adfe81b67dd755cecc-Paper.pdf.

Angelo Mele. A structural model of dense network formation. *Econometrica*, 85(3):825–850, 2017. doi: https://doi.org/10.3982/ECTA10400. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10400.

Angelo Mele, Lingxin Hao, Joshua Cape, and Carey E Priebe. Spectral estimation of large stochastic blockmodels with discrete nodal covariates. *Journal of Business & Economic Statistics*, 41(4):1364–1376, 2022. doi: 10.1080/07350015.2022.2139709.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

Whitney K. Newey and Richard J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004. doi: https://doi.org/10.1111/j.1468-0262.2004.00482.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2004.00482.x`.

J. Pfanzagl and W. Wefelmeyer. A third-order optimum property of the maximum likelihood estimator. *Journal of Multivariate Analysis*, 8(1):1–29, 1978. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(78)90016-7. URL `https://www.sciencedirect.com/science/article/pii/0047259X78900167`.

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. doi: 10.1080/01621459.1971.10482356. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356`.

Paul Rilstone, V.K. Srivastava, and Aman Ullah. The second-order bias and mean squared error of nonlinear estimators. *Journal of Econometrics*, 75(2):369–395, 1996. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(96)89457-7. URL `https://www.sciencedirect.com/science/article/pii/0304407696894577`.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. ISSN 00034851. URL `http://www.jstor.org/stable/2236626`.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878 – 1915, 2011. doi: 10.1214/11-AOS887. URL `https://doi.org/10.1214/11-AOS887`.

Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E. Priebe. A statistical interpretation of spectral embedding: The generalised random dot product graph. *J. R. Stat. Soc., B: Stat. Methodol.*, 84(4):1446–1473, 2022. doi: https://doi.org/10.1111/rssb.12509. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12509`.

Purnamrita Sarkar and Peter J. Bickel. Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962 – 990, 2015. doi: 10.1214/14-AOS1285. URL `https://doi.org/10.1214/14-AOS1285`.

Susanne M. Schennach. Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35(2):634 – 672, 2007. doi: 10.1214/009053606000001208. URL `https://doi.org/10.1214/009053606000001208`.

Srijan Sengupta and Yuguo Chen. A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):365–386, 2018. doi: https://doi.org/10.1111/rssb.12245. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12245`.

Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, and Carey E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012. doi: 10.1080/01621459.2012.699795. URL `https://doi.org/10.1080/01621459.2012.699795`.

Daniel L. Sussman, Minh Tang, and Carey E. Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):48–57, 2014. doi: 10.1109/TPAMI.2013.135.

Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 12 2018. ISSN 0006-3444. doi: 10.1093/biomet/asy054. URL `https://doi.org/10.1093/biomet/asy054`.

Nicholas Syring and Ryan Martin. Gibbs posterior concentration rates under subexponential type losses. *Bernoulli, accepted for publication*, 2022.

Minh Tang and Carey E. Priebe. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415, 2018. ISSN 00905364, 21688966. URL `https://www.jstor.org/stable/26542867`.

Minh Tang, Daniel L. Sussman, and Carey E. Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, 2013. ISSN 00905364, 21688966. URL `http://www.jstor.org/stable/23566864`.

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, Youngser Park, and Carey E. Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2017a. doi: 10.1080/10618600.2016.1193505. URL `https://doi.org/10.1080/10618600.2016.1193505`.

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, and Carey E. Priebe. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23(3):1599 – 1630, 2017b. doi: 10.3150/15-BEJ789. URL `https://doi.org/10.3150/15-BEJ789`.

Minh Tang, Joshua Cape, and Carey E. Priebe. Asymptotically efficient estimators for stochastic blockmodels: The naive MLE, the rank-constrained MLE, and the spectral estimator. *Bernoulli*, 28(2):1049 – 1073, 2022. doi: 10.3150/21-BEJ1376. URL `https://doi.org/10.3150/21-BEJ1376`.

Runze Tang, Michael Ketcha, Alexandra Badea, Evan D. Calabrese, Daniel S. Margulies, Joshua T. Vogelstein, Carey E. Priebe, and Daniel L. Sussman. Connectome smoothing via low-rank approximations. *IEEE Transactions on Medical Imaging*, 38(6):1446–1456, 2019. doi: 10.1109/TMI.2018.2885968.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Fangzheng Xie. Entrywise limit theorems for eigenvectors of signal-plus-noise matrix models with weak signals. *Bernoulli*, 30(1):388–418, 2024. doi: 10.3150/23-BEJ1602. URL `https://doi.org/10.3150/23-BEJ1602`.

Fangzheng Xie and Dingbo Wu. An eigenvector-assisted estimation framework for signal-plus-noise matrix models. *Biometrika*, 111(2):661–676, 2024. doi: https://doi.org/10.1093/biomet/asad058.

Fangzheng Xie and Yanxun Xu. Optimal Bayesian estimation for random dot product graphs. *Biometrika*, 107(4):875–889, 07 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa031. URL https://doi.org/10.1093/biomet/asaa031.

Fangzheng Xie and Yanxun Xu. Efficient estimation for random dot product graphs via a one-step procedure. *Journal of the American Statistical Association*, 118(541):651–664, 2023. doi: 10.1080/01621459.2021.1948419. URL https://doi.org/10.1080/01621459.2021.1948419.

Fangzheng Xie and Yichi Zhang. Higher-order entrywise eigenvectors analysis of low-rank random matrices: Bias correction, Edgeworth expansion and bootstrap. *The Annals of Statistics*, 53(4):1667–1696, 2025.

Stephen J. Young and Edward R. Scheinerman. Random dot product graph models for social networks. In Anthony Bonato and Fan R. K. Chung, editors, *Algorithms and Models for the Web-Graph*, pages 138–149, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-77004-6.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012. doi: 10.1214/12-AOS1036. URL https://doi.org/10.1214/12-AOS1036.