Journal of Machine Learning Research 22 (2021) 1-59 Submitted 10/19; Revised 1/21; Published 10/21

Limit theorems for out-of-sample extensions of the adjacency and Laplacian spectral embeddings

Keith D. Levin KDLEVIN@WISC.EDU
Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706, USA

Fred Roosta FRED.ROOSTA@UQ.EDU.AU
School of Mathematics and Physics, University of Queensland, Brisbane, Australia
International Computer Science Institute, Berkeley, CA 94704, USA

Minh Tang MTANG8@NCSU.EDU
Department of Statistics, North Carolina State University, Raleigh, NC 27696 USA

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
Department of Statistics, University of California at Berkeley, Berkeley, CA 94720, USA
International Computer Science Institute, Berkeley, CA 94704, USA

Carey E. Priebe CEP@JHU.EDU
Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA

Editor: Tina Eliassi-Rad

Abstract

Graph embeddings, a class of dimensionality reduction techniques designed for relational data, have proven useful in exploring and modeling network structure. Most dimensionality reduction methods allow out-of-sample extensions, by which an embedding can be applied to observations not present in the training set. Applied to graphs, the out-of-sample extension problem concerns how to compute the embedding of a vertex that is added to the graph after an embedding has already been computed. In this paper, we consider the out-of-sample extension problem for two graph embedding procedures: the adjacency spectral embedding and the Laplacian spectral embedding. In both cases, we prove that when the underlying graph is generated according to a latent space model called the random dot product graph, which includes the popular stochastic block model as a special case, an out-of-sample extension based on a least-squares objective obeys a central limit theorem. In addition, we prove a concentration inequality for the out-of-sample extension of the adjacency spectral embedding based on a maximum-likelihood objective. Our results also yield a convenient framework in which to analyze trade-offs between estimation accuracy and computational expenses, which we explore briefly. Finally, we explore the performance of these out-of-sample extensions as applied to both simulated and real-world data. We observe significant computational savings with minimal losses to the quality of the learned embeddings, in keeping with our theoretical results.

Keywords: Adjacency Matrix, Spectral Embedding, Graph Laplacian, Out-of-Sample Extension, Random Dot Product Graph

1. Introduction

Graph embeddings are a class of dimensionality reduction techniques designed for network data, which have emerged as a popular tool for exploring and modeling network structure. Given a graph $G = (V, E)$ on vertex set $V = \{1, 2, \ldots, n\}$ with adjacency matrix $A \in \{0, 1\}^{n \times n}$, the graph embedding problem concerns how best to map V to a d-dimensional vector space so that geometry in that vector space captures the topology of G. For example, we may ask that vertices that play similar structural roles in G be mapped to nearby points. Two common approaches to graph embedding are the graph Laplacian embedding (Belkin and Niyogi, 2003; Coifman and Lafon, 2006) and the adjacency spectral embedding (ASE, Sussman et al., 2012), both of which are based on spectral decompositions of the adjacency matrix or a transformation thereof. In many settings, data collection or computational constraints may dictate that having computed an embedding of the graph G, a practitioner may wish to add vertices to G, and compute the corresponding embeddings of these new vertices. We call these new vertices *out-of-sample* vertices, in contrast to the *in-sample* vertices in V. Since constructing the in-sample embedding typically requires a comparatively expensive eigenvalue computation, it is preferable to compute this out-of-sample embedding without computing a new graph embedding from scratch. This problem is well-studied in the dimensionality reduction literature, where it is known as the out-of-sample extension problem. The focus of the present paper is to derive out-of-sample extensions for the ASE and a slight variant of Laplacian eigenmaps, and to establish their statistical properties under a particular natural choice of network model.

Latent space network models are a class of statistical models for graphs in which unobserved geometry drives network formation. Each vertex is assigned a *latent position*, and pairs of vertices form edges according to how near their latent positions are to one another. Under certain latent space models, graph embeddings may be thought of as estimating these latent positions. The focus of the present work is the random dot product graph, a latent position model that subsumes the popular stochastic block model (see Section 1.1 below). Under this model, both the ASE and a slight variant of Laplacian eigenmaps called the Laplacian spectral embedding (LSE; Tang and Priebe, 2018), recover all the latent positions of the in-sample vertices uniformly (Lyzinski et al., 2014; Tang and Priebe, 2018). Specifically, one obtains a bound on the estimation error of order $n^{-1/2}$ (ignoring logarithmic factors) that holds uniformly over all n vertices in the graph. Further, any constant number of vertices jointly obey a CLT, in that their embeddings are jointly asymptotically normally distributed about the true latent positions (Athreya et al., 2016; Levin et al., 2017; Tang and Priebe, 2018). In this paper, we show that analogous results hold for the out-of-sample extensions of both the ASE and LSE. That is, the out-of-sample extensions of these two methods recover the latent positions of the out-of-sample vertices at the same rate as would be obtained by the computationally more expensive in-sample embedding.

1.1 Background and Notation

Most dimensionality reduction and embedding techniques begin with a collection of training data observations $\mathcal{D} = \{z_1, z_2, \ldots, z_n\} \subseteq \mathcal{X}$, where $\mathcal{X}$ is the set of all possible observations (e.g., the set of all possible images, audio signals, etc.). $\mathcal{X}$ is endowed with a similarity measure $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, and most embedding procedures leverage the eigenstructure

of the symmetric similarity matrix $M = [K(z_i, z_j)] \in \mathbb{R}^{n \times n}$. An embedding of the data $\mathcal{D}$ assigns to each $z_i \in \mathcal{D}$ a vector $x_i \in \mathbb{R}^d$, where d is the embedding dimension, with the embeddings $\{x_1, x_2, \ldots, x_n\}$ chosen so as to preserve the structure of the sample $\mathcal{D}$ as captured by the matrix M. This typically manifests as attempting to ensure that elements $z_i, z_j \in \mathcal{D}$ for which $K(z_i, z_j)$ is large are mapped so that $\|x_i - x_j\|$ is small. Suppose that, having computed $x_1, x_2, \ldots, x_n$, we obtain a new *out-of-sample* observation $z \in \mathcal{X}$ (which may or may not appear in the training sample $\mathcal{D}$), which we would like to embed along with the *in-sample* observations $\mathcal{D}$. Letting $\tilde{\mathcal{D}} = \mathcal{D} \cup \{z\}$, a naïve approach would simply construct a new embedding $\{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n, \tilde{x}_{n+1}\}$ based on the sample $\tilde{\mathcal{D}}$. This would involve computational complexity of the same order as that required to compute the initial embedding $\{x_1, x_2, \ldots, x_n\}$. Since computing the embedding $\{x_1, x_2, \ldots, x_n\}$ tends to involve expensive computations, most commonly eigendecompositions, it would be preferable to avoid paying this computational cost repeatedly, particularly if there exists a scheme whereby the embedding $\tilde{x}_{n+1}$ of out-of-sample observation z can be well approximated by a less costly computation. This is the motivation for the out-of-sample (OOS) extension problem, which concerns how to embed z into the same embedding space $\mathbb{R}^d$ based only on the existing *in-sample embedding* $\{x_1, x_2, \ldots, x_n\}$ and the similarity measurements $\{K(z, x_i) : i = 1, 2, \ldots, n\}$. That is, we wish to compute an embedding of z *without* making recourse to the full similarity matrix $M \in \mathbb{R}^{n \times n}$.

As an illustrative example, consider the Laplacian eigenmaps embedding (Belkin and Niyogi, 2003; Belkin et al., 2006). Recall that the *normalized Laplacian* of graph $G = (V, E)$ with adjacency matrix $A \in \mathbb{R}^{n \times n}$ is given by the matrix $L = D^{-1/2} A D^{-1/2}$, where $D \in \mathbb{R}^{n \times n}$ is the diagonal matrix of degrees, with $D_{ii} = \sum_{j=1}^{n} A_{ij}$, and $0^{-1/2} = 0$ by convention (Chung, 1997; Luxburg, 2007; Vishnoi, 2013). The d-dimensional normalized Laplacian eigenmaps embedding of G is then given by the rows of the matrix $\tilde{U} \in \mathbb{R}^{n \times d}$, where the columns of $\tilde{U}$ are the orthonormal eigenvectors corresponding to the top d eigenvalues of L, excluding the trivial eigenvalue 1. Suppose now that we wish to add a vertex v to the graph, to form graph $\tilde{G}$ with adjacency matrix

$$\tilde{A} = \begin{bmatrix} A & \vec{a} \\ \vec{a}^T & 0 \end{bmatrix}, \tag{1}$$

where $\vec{a} \in \{0, 1\}^n$ and has $a_i = 1$ if and only if v forms and edge with in-sample vertex $i \in [n]$. Naïvely, one could simply apply the Laplacian eigenmaps embedding again to $\tilde{A}$, at the cost of another eigendecomposition. Cheaper, however, would be an OOS extension, such as that given by Bengio et al. (2004) or Belkin et al. (2006), that only makes use of the embedding $\tilde{U}$ and the vector of edges $\vec{a}$.

Out-of-sample extensions for multidimensional scaling (MDS, Torgerson, 1952; Borg and Groenen, 2005), spectral clustering (Weiss, 1999; Ng et al., 2002), Laplacian eigenmaps (Belkin and Niyogi, 2003) and ISOMAP (Tenenbaum et al., 2000) appear in Bengio et al. (2004). These extensions were obtained by formulating each of the dimensionality reduction techniques as a least-squares problem, which is possible owing to the fact that the in-sample embeddings are functions of the eigenvalues and eigenvectors of a similarity or distance matrix. Let matrix $M = [K(x_i, x_j)]_{i,j=1}^n$ be the similarity matrix for some similarity function K, and let $\{(\lambda_i, u_i)\}_{i=1}^n$ be the eigenvalue-eigenvector pairs of M. Bengio et al. (2004) derive

the OOS extensions for a number of embeddings as solutions to the least-squares problem

$$\min_{f(x)\in\mathbb{R}^d} \sum_{i=1}^{n} \left(K(x, x_i) - \frac{1}{n} \sum_{j=1}^{d} \lambda_j f_j(x_i) f_j(x) \right)^2,$$

where $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ are the in-sample observations and $f_j(x_i)$ is the i-th component of u_j. A different OOS extension for MDS was considered in Trosset and Priebe (2008). Instead of the least-squares framework of Bengio et al. (2004), Trosset and Priebe (2008) frame the MDS OOS extension problem as a modification of the optimization problem solved by the in-sample MDS embedding.

An approach to the Laplacian eigenmaps OOS extension, different from the one presented here, was pursued in Belkin et al. (2006), incorporating regularization in both the geometry of the training data and the geometry of the similarity function K. Their approach can also be extended to regularized least squares, SVMs, and a variant of SVM in which a Laplacian penalty term is added to the SVM objective. The authors showed that all of these OOS extensions are the solutions to generalized eigenvalue problems. Levin et al. (2015) provides an illustrative example of the practical application of these OOS extensions, using the OOS extension of Belkin et al. (2006) to build an audio search system. More recent OOS extension techniques have attempted to avoid altogether the need to solve least squares or eigenvalue problems, instead training a neural net to learn the embedding, so that at out-of-sample embedding time one need only feed the out-of-sample observation as input to the neural net (see, for example, Quispe et al., 2016; Jansen et al., 2017).

As far as we are aware, the only work to date on the OOS extension for ASE appears in Tang et al. (2013a), in which the authors considered the OOS extension problem for *latent space* models of graphs (see, for example Hoff et al., 2002). These are models in which each vertex has an associated latent vector with edge probabilities given by inner products of the latent vectors. The authors presented an OOS extension based on a least-squares objective and proved a result, analogous to our Theorem 7, given the rate of growth of the error between this out-of-sample embedding and the true out-of-sample latent position. Theorem 7 yields a simplification of the proof of the result originally appearing in Tang et al. (2013a), specialized to the random dot product graph model (see Definition 3 below). We note, however, that our results can be extended to more general latent space network models under suitable conditions on the inner product.

Largely missing from the literature, but of particular importance to the assessment of OOS extensions, is the comparison of the OOS estimate's accuracy compared to its in-sample counter-part. That is, for training sample $\mathcal{D}$ and out-of-sample observation $z \in \mathcal{X}$ (both drawn, perhaps, from a probability distribution on $\mathcal{X}$), how closely does the out-of-sample embedding approximate its in-sample counterpart computed based on $\tilde{\mathcal{D}} = \mathcal{D} \cup \{z\}$? In this work, we address this question as it pertains to the adjacency spectral embedding (ASE) and the Laplacian spectral embedding (LSE; an embedding closely related to the Laplacian eigenmaps embedding but more amenable to analysis; see Section 2). In particular, we show the following:

- Two different approaches to the ASE OOS extension problem yield OOS extensions that recover the true out-of-sample latent position at a rate that matches the in-sample

estimation error rate. The first (Theorem 7), based on a linear least squares objective, holds under essentially no conditions on the model. The second (Theorem 8), based on a maximum-likelihood objective, requires mild regularity conditions.

- An LSE OOS extension based on a linear least-squares objective that, similarly to the ASE OOS extensions, recovers the true out-of-sample latent position at the same rate as the in-sample embedding (Theorem 9).
- Both of the LLS-based OOS extensions obey central limit theorems (Theorems 11 and 13), with each OOS extension asymptotically normally distributed about the true latent position (in the case of ASE) or a transformation thereof (in the case of LSE).

We believe that analogous central limit theorems can be obtained for other OOS extensions such as those presented in Bengio et al. (2004) and for the maximum-likelihood ASE OOS extension, but do not pursue this generalization here. We note that the ASE out-of-sample extensions analyzed here were first presented and analyzed in Levin et al. (2018). This work extends that conference paper by giving complete proofs, extending the analysis to the Laplacian spectral embedding (requiring substantially more involved proofs), and adding more thorough experimental results.

1.2 Notation

Before continuing, we pause to establish notation. For a matrix $M \in \mathbb{R}^{n_1 \times n_2}$, we denote by $\sigma_i(M)$ the i-th singular value of M, so that $\sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq \sigma_k(M) \geq 0$, where $k = \min\{n_1, n_2\}$. For integer $k > 0$, we let $[k] = \{1, 2, \ldots, n\}$. Throughout the paper, n will denote the number of vertices in the observed graph G. For a vector x, the unadorned norm $\|x\|$ will denote the Euclidean norm of x, while for all $p > 0$, $\|x\|_p$ will denote the p-norm of x, where $\|x\|_\infty = \max_i |x_i|$. For a matrix M, $\|M\|_F$ will denote the Frobenius norm, $\|M\|$ will denote the spectral norm

$$\|M\| = \sup_{x:\|x\|=1} \|Mx\|,$$

and $\|M\|_{2,\infty}$ will denote the 2-to-∞ norm,

$$\|M\|_{2,\infty} = \sup_{x:\|x\|=1} \|Mx\|_\infty.$$

Most of our results will concern the behavior of certain quantities as the number of vertices n increases to ∞. We will often, for ease of notation, suppress this dependence on n, but it should be assumed throughout that all quantities are dependent on n, with the exception of the latent position distribution F and the latent space dimension d (see Definition 3). Thus, for example, we will in several places refer to a "sequence of matrices" $Q \in \mathbb{R}^{d \times d}$, where we suppress what ought to be, say, a subscript n. Throughout, $C > 0$ denotes a positive constant, not depending on n, whose value may change from line to line or even, occasionally, within the same line. Given an event E, we let E^c denote its complement, and let $\Pr[E]$ denote the probability of event E (the probability measure in question will always be clear from context). Given a collection of events $\{E_n\}$ indexed by

n, suppose that with probability 1 there exists n_0 such that E_n occurs whenever $n \geq n_0$. If this is the case, we say that E_n occurs eventually or, by a slight abuse of terminology, say simply that E_n occurs.

We make standard use of the big-O, big-Ω and big-Θ notation. Thus, for example, we write $f(n) = O(g(n))$ to denote the existence of a constant $C > 0$ such that for all suitably large n, $f(n) \leq Cg(n)$. We write $f(n) = \tilde{O}(g(n))$ to mean that $f(n) = O(g(n))$ ignoring logarithmic factors. That is, if there exists a $c > 0$ such that $f(n) = O(g(n) \log^c n)$ (throughout the paper, c is never larger than 2 or 3 and is typically 1/2). Our one slight abuse of this notation is in the case where, letting $\{Z_n\}$ be a sequence of random variables, we write $Z_n = O(g(n))$ to mean that there exists a constant $C > 0$ such that almost surely there exists n_0 such that $|Z_n| \leq Cg(n)$ for all $n \geq n_0$, replacing the modulus with an appropriate norm when Z_n is a vector or matrix. The concentration inequalities in the sequel are all of this form. We note that throughout, we prove these results by showing first that $\Pr[|Z_n| \geq Cg(n)] \leq Cn^{-(1+\epsilon)}$ is summable for all suitably small $\epsilon > 0$. We then use the independence of $\{Z_n : n = 1, 2, \dots\}$ to invoke the Borel-Cantelli lemma (Billingsley, 1995) to conclude that $Z_n = O(g(n))$. Thus, though these concentration inequalities are stated as holding asymptotically, they all have finite-sample analogues obtained in the course of their proofs.

1.3 Roadmap

The remainder of this paper is structured as follows. In Section 2, we formalize the graph out-of-sample extension problem, and introduce a few methods for constructing such extensions. In Section 3, we present our main theoretical results, proving concentration and asymptotic distributions for these extensions. Section 4 gives an experimental investigation of the properties of these embeddings. We conclude in Section 5 with a brief discussion of directions for future work.

2. Out-of-sample Extension for ASE and LSE

Given a graph $G = ([n], E)$ with adjacency matrix $A \in \{0, 1\}^{n \times n}$, the adjacency spectral embedding (ASE; Sussman et al., 2012) and the Laplacian spectral embedding (LSE; Tang and Priebe, 2018) each provide a mapping of the n vertices of G into $\mathbb{R}^d$. The ASE maps the vertices of G to d-dimensional representations $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n \in \mathbb{R}^d$ given by the rows of the matrix

$$\hat{X} = \mathrm{ASE}(A, d) = \hat{U}\hat{S}^{1/2} \in \mathbb{R}^{n \times d}, \tag{2}$$

where $\hat{S} \in \mathbb{R}^{d \times d}$ is the diagonal matrix with entries given by the top d eigenvalues of A and the columns of $\hat{U} \in \mathbb{R}^{n \times d}$ are the corresponding orthonormal eigenvectors. The Laplacian spectral embedding (LSE; Tang and Priebe, 2018) proceeds according to a similar eigenvalue truncation, applied to the normalized graph Laplacian,

$$L = \mathcal{L}(A) := D^{-1/2} A D^{-1/2},$$

where $D \in \mathbb{R}^{n \times n}$ is the diagonal degree matrix, with $D_{i,i} = \sum_{j=1}^{n} A_{i,j}$, with $0^{-1/2} = 0$ by convention. The LSE embeds the vertices of G as $\check{X}_1, \check{X}_2, \dots \check{X}_n \in \mathbb{R}^d$ given by the rows of

the matrix

$$\check{X} = \mathrm{LSE}(A, d) = \check{U}\check{S}^{1/2} \in \mathbb{R}^{n \times d}, \tag{3}$$

where $\check{S} \in \mathbb{R}^{d \times d}$ is the diagonal matrix formed of the d largest-magnitude eigenvalues of the graph Laplacian L and $\check{U} \in \mathbb{R}^{n \times d}$ is the matrix formed of the d corresponding orthonormal eigenvectors. The well-known Laplacian eigenmaps embedding (Belkin and Niyogi, 2003) is based on the eigenvectors corresponding to the smallest eigenvalues of the matrix $I - D^{-1/2}AD^{-1/2} = I - L$. Since this matrix has the same eigenspace as L, with the eigenvalue ordering reversed, the Laplacian eigenmaps embedding is given by the rows of $\hat{U} \in \mathbb{R}^{n \times d}$. Thus, results similar to those presented here for the LSE can be obtained for the Laplacian eigenmaps embedding, as well.

We note that in both of the embeddings just described, there may be a concern that the d largest-magnitude eigenvalues need not all be positive, and hence square roots $\hat{S}^{1/2}$ and $\check{S}^{1/2}$ will be ill-defined. As a result, it may be preferable, in general, to consider instead the top-d singular values of A and L. We will not consider this issue in the present work, since under the model considered in this paper, with probability 1, the d largest-magnitude eigenvalues will be positive for all suitably large n (see Definition 3 below and Lemma 14 in Appendix A).

Remark 1 (Comparing ASE and LSE) *Both the ASE and LSE yield low-dimensional representations of the vertices of G, and it is natural to ask which embedding is preferable. The answer, in general, is dependent on the precise model under consideration and the intended downstream task. For example, one can show that neither the ASE nor the Laplacian embedding strictly dominates in a vertex clustering task. Section 4 of Tang and Priebe (2018) demonstrates that ASE performs better than the Laplacian embedding when applied to graphs with a core-periphery structure. Such structures are ubiquitous in real networks; see, for example, Leskovec et al. (2009) and Jeub et al. (2015). We refer the interested reader to Cape et al. (2019) for a more thorough theoretical treatment of this point. The differing behaviors of the ASE and LSE can be related to the emerging distinction in the literature between node embeddings, in which one seeks to preserve closeness in the topology of the graph, and structural graph representations, in which one seeks to embed vertices near one another according to the similarity of their structural roles in the graph (Srinivasan and Ribero, 2020). These two views correspond roughly to the LSE and ASE, respectively.*

The two embeddings just discussed are especially well-suited to the random dot product graph (RDPG; Young and Scheinerman, 2007; Athreya et al., 2018), a model in which graph structure is driven by the geometry of latent positions associated to the vertices.

Definition 2 (Inner product distribution) *A distribution F on $\mathbb{R}^d$ is a* d-dimensional inner product distribution *if $0 \leq x^T y \leq 1$ whenever $x, y \in \operatorname{supp} F$.*

Definition 3 (Random Dot Product Graph) *Let F be a d-dimensional inner product distribution, and let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F$ be collected in the rows of $X \in \mathbb{R}^{n \times d}$. Let G be a random graph with adjacency matrix $A \in \{0,1\}^{n \times n}$. We say that G is a* random dot product graph *(RDPG) with* latent positions *$X_1, X_2, \ldots, X_n \in \mathbb{R}^d$, if the edges of G are*

independent conditioned on $\{X_1, X_2, \ldots, X_n\}$*, with*

$$\Pr[A|X] = \prod_{1 \leq i < j \leq n} (X_i^T X_j)^{A_{i,j}} (1 - X_i^T X_j)^{1 - A_{i,j}}. \tag{4}$$

We say that X_i *is the latent position associated to the* i*-th vertex in* G*, and write* $(A, X) \sim \mathrm{RDPG}(F, n)$ *to mean that the rows of* $X \in \mathbb{R}^{n \times d}$ *are drawn i.i.d. from* F *and that* $A \in \{0,1\}^{n \times n}$ *is generated according to Equation* (4) *conditional on* X*.*

Note that the RDPG has an inherent nonidentifiability, owing to the fact that the distribution of A is unchanged by an orthogonal rotation of the latent positions: for latent position matrix $X \in \mathbb{R}^{n \times d}$ and orthogonal matrix $W \in \mathbb{R}^{d \times d}$, both $X \in \mathbb{R}^{n \times d}$ and $XW \in \mathbb{R}^{n \times d}$ give rise to the same distribution over adjacency matrices, in that $\mathbb{E}[A \mid X] = XX^T = XW(XW)^T$. Thus, we can only ever hope to recover the latent positions of the RDPG up to some orthogonal transformation. Throughout this work, we denote by $\Delta = \mathbb{E}X_1 X_1^T \in \mathbb{R}^{d \times d}$ the second moment matrix of the latent position distribution F. Our results require that Δ be of full rank, an assumption that we make without loss of generality owing to the fact that if Δ is of, say, rank $d' < d$, then we may equivalently think of F as a d'-dimensional inner product distribution by restricting our attention to an appropriate d'-dimensional subspace of $\mathbb{R}^d$.

Remark 4 (Extension to other graph models) *As alluded to above, the RDPG as defined here only captures graphs with positive semi-definite expected adjacency matrices. This limitation can be avoided by considering the* generalized *RDPG (Rubin-Delanchy et al., 2017). The results stated in the present work can for the most part be extended to this model, at the expense of additional notational complexity, which we prefer to avoid here. Similarly, using standard concentration inequalities, most of the results presented here can be extended beyond binary edges to consider independent edges that are unbiased (*$\mathbb{E}A_{i,j} = X_i^T X_j$*) with sub-Gaussian or sub-gamma tails (Boucheron et al., 2013; Tropp, 2015).*

Remark 5 (Incorporating Sparsity) *As defined above, the RDPG produces only dense graphs. That is, the number of edges grow quadratically in the number of vertices:* $\sum_{i<j} A_{ij} = \Omega(n^2)$*. This behavior is in contrast with many real-world networks, which are* sparse*, in the sense that the number of edges that are present in the network is much smaller than the* $O(n^2)$ *possible edges. A simple way to introduce this sparse behavior is to shrink the latent positions toward the origin as* n *increases, replacing the latent position matrix* X *with* $\sqrt{\rho_n} X$*, where* $\rho_n \in [0, 1]$ *is a* sparsity parameter*. Then* $n^{-2} \sum_{i<j} A_{ij} = \Theta(\rho_n)$*. Taking* $\rho_n \to 0$ *as* $n \to \infty$ *imposes sparsity, in the sense that the number of edges now grows more slowly than the* $O(n^2)$ *rate predicted by the (unscaled) RDPG.*

For ease of notation and exposition, we ignore sparsity in the material below, assuming throughout that F *is fixed in* n*. Nonetheless, the results presented below continue to hold with minor modification in the presence of a sparsity parameter* ρ_n*. Roughly speaking, so long as* $n\rho_n = \omega(\log^c n)$ *for a suitably large constant* $c > 0$*, then the results presented in this paper continue to hold once we replace each factor of* n *with* $\rho_n n$*. We refer the interested reader to Tang and Priebe (2018); Rubin-Delanchy et al. (2017); Levin et al. (2019) for examples of similarly-motivated concentration inequalities and central limit theorems that incorporate sparsity.*

Throughout this paper, we will assume that $(A, X) \sim \text{RDPG}(F, n)$ for some d-dimensional inner product distribution F, and write $P = \mathbb{E}[A \mid X] = XX^T$. Then $\hat{X} = \text{ASE}(A, d)$ is a natural alternate estimate of the matrix of true latent positions X. Similarly, $\check{X} = \text{LSE}(A, d)$ is a natural estimate of $\tilde{X} = T^{-1/2}X$, where $T \in \mathbb{R}^{n \times n}$ is a diagonal matrix with entries $T_{i,i} = \sum_j X_j^T X_i$. The rows of $\tilde{X}$ can be thought of as the Laplacian spectral embeddings of the matrix $P = XX^T$, in the sense that $\tilde{X}\tilde{X}^T = \mathcal{L}(P)$. It has been shown previously that the ASE consistently estimates the latent positions in the RDPG (Sussman et al., 2012; Tang et al., 2013b), and successfully recovers community structure in the (positive semi-definite) stochastic block model (Lyzinski et al., 2014), which can be recovered as a special case of the RDPG by taking the distribution F to be a mixture of point masses. Similar results can be shown for the LSE (Tang and Priebe, 2018).

Lemma 6 *Let $(A, X) \sim \text{RDPG}(F, n)$ for some d-dimensional inner product distribution F and let $\hat{X}, \check{X}, \tilde{X} \in \mathbb{R}^{n \times d}$ be as above. Then there exists a sequence of orthogonal matrices $Q \in \mathbb{R}^{d \times d}$ such that*

$$\|\hat{X} - XQ\|_{2,\infty} = O\left(\frac{\log n}{\sqrt{n}}\right). \tag{5}$$

Further, if there exists a constant $\eta > 0$ such that $\eta \leq x^T y \leq 1 - \eta$ whenever $x, y \in \operatorname{supp} F$, then there exists a sequence of orthogonal matrices $\tilde{Q} \in \mathbb{R}^{d \times d}$ such that

$$\|\check{X} - \tilde{X}\tilde{Q}\|_{2,\infty} = O\left(\frac{\log^{1/2} n}{n}\right). \tag{6}$$

Proof The bound in Equation (5) is Lemma 5 in Lyzinski et al. (2014). Equation (6) follows by a broadly similar argument, once one accounts for the additional randomness from the vertex degrees in the Laplacian. The Laplacian result requires the additional assumption surrounding $\eta > 0$ to account for the fact that we are normalizing by the degrees (i.e., dividing by $X_i^T \mu$). Details can be found in Appendix A. ■

The reader may note that the rates obtained in Lemma 6 for estimating X and $\tilde{X}$ differ by a factor of $\sqrt{n}$. This difference is due to the fact that $\tilde{X} = T^{-1/2}X$ is obtained by rescaling the rows of X by the square roots of the (expected) vertex degrees, which shrinks them toward the origin. As a result, $\hat{X}$ and $\tilde{X}$ will require different rescalings to ensure nondegeneracy in our large-n asymptotic results below.

As an aside, one might ask that we estimate X or $T^{-1/2}X$ according to maximum-likelihood, instead of least-squares. Unfortunately, maximum-likelihood estimation of X (or a transformation thereof) is impractical, if not altogether intractable. However, one can show that in the special case of the stochastic blockmodel, the ASE is asymptotically efficient, in that it recovers the true model parameters at a rate that matches the maximum-likelihood estimate (Tang et al., 2017). We are not aware of an analogous result for the more general RDPG, but recent minimax results indicate that the ASE recovers X at a rate that is optimal up to log factors (Xie and Xu, 2020), and is thus (nearly) asymptotically efficient.

Suppose that a graph $G = ([n], E)$ with adjacency matrix $A \in \mathbb{R}^{n\times n}$ is a random dot product graph, so that $(A, X) \sim \mathrm{RDPG}(F, n)$, and we compute

$$\hat{X} = \mathrm{ASE}(A, d) = [\hat{X}_1 \hat{X}_2 \cdots \hat{X}_n]^T \in \mathbb{R}^{n\times d} \text{ and } \check{X} = \mathrm{LSE}(A, d) = [\check{X}_1 \check{X}_2 \cdots \check{X}_n]^T \in \mathbb{R}^{n\times d},$$

where $\hat{X}_i, \check{X}_i \in \mathbb{R}^d$ are embeddings of the i-th vertex under ASE and LSE, respectively. Suppose now that a vertex v having latent position $\bar{w} \in \operatorname{supp} F$ is added to the graph G to form $\tilde{G} = ([n] \cup \{v\}, E \cup E_v)$, where $E_v \subseteq \{\{i, v\} : i = 1, 2, \ldots, n\}$. The edges between the out-of-sample vertex v and the in-sample vertices $\{1, 2, \ldots, n\}$ are specified by a vector $\vec{a} \in \{0, 1\}^n$ such that $a_i = 1$ if $\{i, v\} \in E_v$ and $a_i = 0$ otherwise. Thus, $\tilde{G}$ has adjacency matrix $\tilde{A}$ as in Equation (1) above. Having computed an embedding $\hat{X}$ or $\check{X}$, we would like to embed the vertex v to obtain an estimate of the true latent position $\bar{w}$ (in the case of ASE) or, in the case of LSE, its Laplacian spectral embedding $\tilde{w} = \bar{w}/\sqrt{n\mu^T\bar{w}} \in \mathbb{R}^d$, where $\mu = \mathbb{E}X_1$ is the mean of F. In the case of ASE, the out-of-sample extension problem concerns how to compute an estimate of $\bar{w}$ based only on $\hat{X}$ and $\vec{a}$. Similarly, in the case of LSE, the out-of-sample extension problem requires computing an estimate of $\tilde{w}$ based only on the information in $\hat{X}$, $\vec{a}$ and, for reasons that will become clear below, the vector of in-sample vertex degrees, $\vec{d} \in \mathbb{R}^n$.

2.1 Out-of-sample extension for ASE

Two natural approaches to the out-of-sample extension of ASE suggest themselves. The first, following Bengio et al. (2004), involves embedding the out-of-sample vertex v as

$$\hat{w}_{\mathrm{LS}} = \arg\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} \left(a_i - \hat{X}_i^T w\right)^2, \tag{7}$$

where a_i is the i-th component of the vector $\vec{a} \in \mathbb{R}^n$ of edges between the out-of-sample vertex and the in-sample vertices. We refer to $\hat{w}_{\mathrm{LS}}$ as the *linear least squares out-of-sample* (LLS OOS) extension of adjacency spectral embedding.

Like the ASE, $\hat{w}_{\mathrm{LS}}$ is the solution to a least-squares problem. As mentioned above, the motivation for defining the ASE as in Equation (2) is that maximum-likelihood estimation of the nd-dimensional X is computationally intractable, in practice. In the case of the out-of-sample extension problem, on the other hand, our goal is to estimate the d-dimensional $\bar{w}$ based on $O(n)$ edges and the in-sample latent position estimates $\hat{X}$, and a maximum-likelihood version of the problem is feasible. Conditional on the true latent positions $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ of the in-sample vertices and the true latent position $\bar{w} \in \mathbb{R}^d$ of the out-of-sample vertex, the entries of $\vec{a}$ are independent Bernoulli random variables, with $a_i \sim \mathrm{Bernoulli}(X_i^T\bar{w})$. Thus, the log likelihood (conditional on the in-sample latent positions) is

$$\ell(w) = \sum_{i=1}^{n} a_i \log X_i^T w + (1 - a_i) \log(1 - X_i^T w).$$

Of course, in practice we observe the latent positions only through their ASE estimates $\{\hat{X}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. Thus, we define the maximum-likelihood out-of-sample extension for ASE

as the maximizer of the plug-in likelihood, i.e., as the solution to

$$\max_{w\in\mathbb{R}^d} \sum_{i=1}^{n} a_i \log \hat{X}_i^T w + (1-a_i)\log\left(1-\hat{X}_i^T w\right). \tag{8}$$

Unfortunately, this objective need not achieve its optimum inside the support of F. Indeed, the objective need not even be bounded. Thus, we will settle for a slight reformulation of this objective, and define the maximum-likelihood out-of-sample (ML OOS) extension for ASE to be the solution to a constrained maximum-likelihood problem,

$$\hat{w}_{\mathrm{ML}} = \arg\max_{w\in\widehat{\mathcal{T}}_\epsilon} \sum_{i=1}^{n} a_i \log \hat{X}_i^T w + (1-a_i)\log\left(1-\hat{X}_i^T w\right), \tag{9}$$

where $\widehat{\mathcal{T}}_\epsilon = \{w \in \mathbb{R}^d : \epsilon \le \hat{X}_i^T w \le 1-\epsilon, i \in [n]\}$, and $\epsilon > 0$ is some small constant. We note that we call this the maximum-likelihood OOS extension, though it is, strictly speaking, based on a plug-in approximation to the true likelihood given in Equation (8).

Note that, as required by the out-of-sample problem, both $\hat{w}_{\mathrm{LS}}$ and $\hat{w}_{\mathrm{ML}}$ are functions only of the in-sample embedding $\hat{X} \in \mathbb{R}^{n\times d}$ and the edges between the out-of-sample vertex v and the in-sample vertices $[n]$, as encoded in the vector $\vec{a} \in \mathbb{R}^n$.

2.2 Out-of-sample extension for LSE

Recall that given the adjacency matrix A of graph $G = ([n], E)$, we form the sample graph Laplacian $L = \mathcal{L}(A) = D^{-1/2}AD^{-1/2}$ and embed in-sample vertex $i \in [n]$ as $\check{X}_i \in \mathbb{R}^d$, the i-th row of

$$\check{X} = \check{U}\check{S}^{1/2} \in \mathbb{R}^{n\times d},$$

where we remind the reader that $\check{U} \in \mathbb{R}^{n\times d}$ denotes the matrix formed by the top d orthonormal eigenvectors of L with their corresponding eigenvalues collected in the diagonal matrix $\check{S} \in \mathbb{R}^{d\times d}$. Conditional on the latent positions $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$, we have $\mathbb{E}[A|X] = XX^T = P \in \mathbb{R}^{n\times n}$, and we view $L = \mathcal{L}(A)$ as an estimate of $\mathcal{L}(P) = T^{-1/2}PT^{-1/2}$, where $T \in \mathbb{R}^{n\times n}$ is the matrix of (conditional) expected degrees, $T_{i,i} = \sum_{j=1}^n P_{i,j} = \sum_{j=1}^n X_i^T X_j$. Applying the LSE to $\mathcal{L}(P)$, we may think of the rows of

$$\tilde{X} = \tilde{U}\tilde{S}^{1/2} \in \mathbb{R}^{n\times d}$$

as the "true" Laplacian spectral embedding, and view $\check{X}$ as an estimate of this quantity.

Given out-of-sample vertex v with latent position $\bar{w} \in \mathbb{R}^d$, the natural Laplacian embedding of v, in light of the definition of $\tilde{X}$, is given by $\tilde{w} = \bar{w}/\sqrt{n\mu^T\bar{w}}$, where $\mu = \mathbb{E}X_1 \in \mathbb{R}^d$ is the mean of F. Of course, in practice we must compute the out-of-sample embedding of v based on $\check{X} \in \mathbb{R}^{n\times d}$ and the vector of edges $\vec{a} \in \mathbb{R}^n$ to obtain an estimate of $\tilde{w}$. In applying the least-squares approach suggested by Equation (7) and used in Bengio et al. (2004), it is most natural to consider the minimizer

$$\check{w}_{\mathrm{LS}} = \arg\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} \left(\frac{a_i}{\sqrt{d_v d_i}} - \check{X}_i^T w\right)^2, \tag{10}$$

where $d_i = \sum_{j=1}^{n} A_{i,j}$ is the degree of the i-th in-sample vertex, and $d_v = \sum_i a_i$ is the degree of the out-of-sample vertex v. We refer to $\check{w}_{\mathrm{LS}}$ as the LLS OOS extension of the Laplacian spectral embedding. We note that Equation (10) requires that we keep in-sample vertex degree information for use in the out-of-sample extension, which violates the typical requirement that we compute the out-of-sample extension using only $\check{X}$ and $\vec{a}$. Nonetheless, it is reasonable to allow the use of the vector $\vec{d}$, since typically the embedding dimension d is of a smaller order than n and thus the space required to store node degrees is of the same or smaller order as that required to store $\check{X} \in \mathbb{R}^{n \times d}$. We note that one could avoid this additional storage by replacing d_i with $\sum_{j=1}^{n} \check{X}_j^T \check{X}_i$ and all our results below would go through (see Lemma 18), but this would come at the expense of notational inconvenience and longer proofs below. The motivation for the least-squares objective in Equation (10) becomes clear if we think of $d_v^{-1/2} d_i^{-1/2} a_i$ as an estimate of the normalized kernel

$$\bar{K}(i, v) = \frac{X_i^T \bar{w}}{n\sqrt{X_i^T \mu \bar{w}^T \mu}},$$

where $\mu \in \mathbb{R}^d$ is again the mean of F.

3. Theoretical Results

The main results of this paper concern concentration inequalities and central limit theorems for the OOS extensions introduced in Section 2. We first present the concentration inequalities, which allow us to control the rate of convergence of the OOS extension to the parameter of interest, given by the true OOS latent position $\bar{w}$ in the case of ASE, and by the transformed latent position $\tilde{w} = \bar{w}/\sqrt{n\mu^T \bar{w}}$ in the case of LSE.

3.1 Rates of convergence for OOS extensions

A first question surrounding the OOS extensions presented in the preceding section concerns their quality as estimators of their respective true parameters. Interestingly, all of the OOS extensions presented above recover their respective target parameters at asymptotic rates that match that of the full-graph embedding.

We begin by considering the ASE OOS extensions defined in Equations (7) and (9). Both of these estimates recover the true out-of-sample latent position $\bar{w}$ at the same asymptotic rate (see Theorems 7 and 8 below), and this rate matches the one we would obtain if we were to compute the ASE of the augmented graph $\tilde{G}$ with adjacency matrix $\tilde{A}$, given in Lemma 6. We find that the estimation error between the least squares OOS extension for ASE $\hat{w}_{\mathrm{LS}}$ and the true latent position $\bar{w}$ follows the same rate.

Theorem 7 *Let F be a d-dimensional inner-product distribution and suppose $(A, X) \sim \mathrm{RDPG}(F, n)$. Let v denote the out-of-sample vertex, with latent position $\bar{w} \in \mathbb{R}^d$ satisfying $0 \leq \bar{w}^T x \leq 1$ for all $x \in \operatorname{supp} F$. Let $\hat{w}_{\mathrm{LS}}$ denote the LS-based OOS extension for ASE based on $\hat{X} = \mathrm{ASE}(A, d)$ and the vector of edges $\vec{a} \in \mathbb{R}^n$ between v and the in-sample vertices, as defined in Equation (7). There exists a sequence of orthogonal matrices $Q \in \mathbb{R}^{d \times d}$ such that*

$$\|Q\hat{w}_{\mathrm{LS}} - \bar{w}\| = O(n^{-1/2} \log n),$$

and this matrix Q is the same one guaranteed by Lemma 6.

Proof By definition in Equation (7), $\hat{w}_{\mathrm{LS}}$ is the solution to a least squares problem that minimizes $\|\hat{X}w - \vec{a}\|$ over all $w \in \mathbb{R}^d$. By Lemma 6, $\hat{X}$ is close to the matrix of true latent positions X. Letting $w_{\mathrm{LS}} \in \mathbb{R}^d$ denote the least squares solution if one uses the true latent positions X in place of $\hat{X}$, Lemma 26 in Appendix B uses a standard result for solutions of perturbed linear systems to show that with high probability, $\|Q\hat{w}_{\mathrm{LS}} - w_{\mathrm{LS}}\| \leq Cn^{-1/2}\log n$, where $Q \in \mathbb{R}^{d\times d}$ is the orthogonal matrix guaranteed by Lemma 6.

Using basic linear algebra, we can bound $\|w_{\mathrm{LS}} - \bar{w}\| \leq C\|X^T(\vec{a} - X\bar{w})\|/\sigma_d(X)$. Lemma 27 in Appendix B uses the fact that the singular values of X grow linearly under the RDPG to lower-bound the denominator and controls the numerator using Hoeffding's inequality to obtain $\|w_{\mathrm{LS}} - \bar{w}\| = O(n^{-1/2}\log n)$. The result then follows by a triangle inequality applied to $\|Q\hat{w}_{\mathrm{LS}} - \bar{w}\|$. A detailed proof can be found in Appendix B. ■

In a similar vein, the ML-based OOS extension also recovers the true out-of-sample latent position at a rate that matches that of the in-sample embedding, given by Equation (5) in Lemma 6.

Theorem 8 *Let F be a d-dimensional inner-product distribution for which there exists a constant $\eta > 0$ such that $\eta < x^T y < 1 - \eta$ for all $x, y \in \operatorname{supp} F$. Suppose that $(A, X) \sim \mathrm{RDPG}(F, n)$ and let v be an out-of-sample vertex with latent position $\bar{w} \in \mathbb{R}^d$ and satisfying $\eta < \bar{w}^T x < 1 - \eta$ for all $x \in \operatorname{supp} F$. Let $\hat{w}_{\mathrm{ML}}$ be the out-of-sample embedding defined in Equation (9), with $\epsilon > 0$ chosen so that $\epsilon < \eta$. Then there exists a sequence of orthogonal matrices $Q \in \mathbb{R}^{d\times d}$ such that*

$$\|Q\hat{w}_{\mathrm{ML}} - \bar{w}\| = O(n^{-1/2}\log n),$$

and this matrix Q is the same one guaranteed by Lemma 6.

Proof Using the definition of $\widehat{\mathcal{T}}_\epsilon$ and a standard argument from convex optimization, Lemma 28 shows that with probability 1, it holds for all suitably large n that

$$\|Q\hat{w}_{\mathrm{ML}} - \bar{w}\| \leq \frac{C\|\nabla\hat{\ell}(Q^T\bar{w})\|}{n}.$$

Lemma 29 uses the triangle inequality and standard concentration inequalities to bound

$$\|\nabla\hat{\ell}(Q^T\bar{w})\| = O(\sqrt{n}\log n),$$

and the result follows by combining the above two displays. A detailed proof can be found in Appendix C. ■

In keeping with the above two results, the least-squares LSE OOS extension given in Equation (10) recovers the true out-of-sample Laplacian embedding $\tilde{w}$ at a rate that matches that of the Laplacian spectral embedding $\tilde{w}$ of the augmented graph $\tilde{G}$, given by Equation (6) in Lemma 6.

Theorem 9 *Let F be a d-dimensional inner-product distribution with mean $\mu = \mathbb{E}X_1$, and suppose that there exists a constant $\eta > 0$ such that $\eta < x^T y < 1 - \eta$ for all $x, y \in \operatorname{supp} F$. Let $(A, X) \sim \mathrm{RDPG}(F, n)$, let v be an out-of-sample vertex with latent position $\bar{w} \in \mathbb{R}^d$ and satisfying $\eta < \bar{w}^T x < 1 - \eta$ for all $x \in \operatorname{supp} F$, and let $\tilde{w} = \bar{w}/\sqrt{n\mu^T \bar{w}}$ be the Laplacian spectral embedding of this latent position. Then there exists a sequence of orthogonal matrices $\tilde{Q} \in \mathbb{R}^{d \times d}$ such that*

$$\|\tilde{Q}\check{w}_{\mathrm{LS}} - \tilde{w}\| \leq Cn^{-1} \log^{1/2} n,$$

and this matrix $\tilde{Q}$ is the same one guaranteed by Lemma 6.

Proof Letting $\tilde{w}_{\mathrm{LS}}$ denote the LLS OOS solution if we had access to the true latent positions, the triangle inequality and unitary invariance of Euclidean norm imply

$$\|\tilde{Q}\check{w}_{\mathrm{LS}} - \tilde{w}\| \leq \|\tilde{Q}\check{w}_{\mathrm{LS}} - \tilde{w}_{\mathrm{LS}}\| + \|\tilde{w}_{\mathrm{LS}} - \tilde{w}\|. \tag{11}$$

Lemma 31 in Appendix D bounds the first right-hand term as $O(n^{-1} \log^{1/2} n)$. The proof relies on a perturbed least squares argument broadly similar to that used in the proof of Theorem 7, though now requiring a more careful argument to account for the renormalization by the degrees in the graph Laplacian.

Lemma 30 in Appendix D bounds

$$\|\tilde{w}_{\mathrm{LS}} - \tilde{w}\| \leq \frac{2\|\tilde{X}^T (d_v^{-1/2} D^{-1/2} \vec{a} - \tilde{X}\tilde{w})\|}{\sigma_d^2(\tilde{X})}, \tag{12}$$

where D is the matrix of in-sample degrees and d_v is the degree of the out-of-sample vertex. Lemma 19 in Appendix A implies that $\sigma_d^2(\tilde{X}) = \Theta(1)$, and basic concentration inequalities bound the numerator of Equation (12) as $O(n^{-1} \log^{1/2} n)$. Thus, the second of the two right-hand terms in Equation (11) also grows at the rate $O(n^{-1} \log^{1/2} n)$. A detailed proof is given in Appendix D. ■

Remark 10 (Adversarial selection of the OOS vertex) *It is natural to ask how the concentration results just described might be adapted to the setting in which an adversary selects the out-of-sample vertex. Such a setup would presumably consist of an adversary selecting the OOS vertex* after *observing the network. Investigating the OOS extension problem under this setting would require analysis* conditional *on the network $\tilde{A}$, which would render several of our technical lemmas inapplicable. We note, however, that arguments similar to those in the proofs of Theorems 7, 8 and 9, show that a network $\tilde{A}$ generated according to to* $\mathrm{RDPG}(F, n+1)$ *is impervious to adversarial selection, in a certain sense. Specifically, one can show that with high probability, there exists no vertex in $\tilde{A}$ whose OOS extension differs from its true latent position (up to rotational nonidentifiability) by more than a factor of $\tilde{O}(n^{-1/2})$ in the case of the ASE, and $\tilde{O}(n^{-1})$ in the case of the LSE. An alternative adversarial model would be one in which the adversary is permitted to choose the latent position $\bar{w}$ of the out-of-sample vertex before the generation of the edges. Since our concentration results require only that $\bar{w}$ be consistent with the RDPG (in that its inner*

products with $\operatorname{supp} F$ *lie in* $[0,1]$*), a slight adaptation of our results above show that the OOS extensions presented in Section 2 are robust to this adversarial model. We leave a more thorough exploration of adversarial variants of the out-of-sample extension problem for future work.*

3.2 Central limit theorems for the OOS extensions

We now turn our attention to the question of the asymptotic distribution of the OOS extensions introduced in Section 2. Once again, we state the results for the case of Bernoulli edges, but similar results can be shown for a broader class of edge noise models, provided that noise model and the latent position distribution F obey suitable moment conditions.

Theorem 11 *Let* F *be a* d*-dimensional inner-product distribution and suppose that* $(A, X) \sim \operatorname{RDPG}(F, n)$ *and let* v *be the out-of-sample vertex with latent position* $\bar{w} \in \mathbb{R}^d$ *satisfying* $0 \leq \bar{w}^T x \leq 1$ *for all* $x \in \operatorname{supp} F$*. Let* $\hat{w}_{\mathrm{LS}}$ *be the least-squares OOS extension as defined in Equation* (7)*. Then there exists a sequence of orthogonal* d*-by-*d *matrices* Q *such that*

$$\sqrt{n}(Q\hat{w}_{\mathrm{LS}} - \bar{w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{F,\bar{w}}),$$

where for any w *satisfying* $0 \leq w^T x \leq 1$ *for all* $x \in \operatorname{supp} F$*, we define*

$$\Sigma_{F,w} = \Delta^{-1}\mathbb{E}\left[X_1^T w(1 - X_1^T w)X_1 X_1^T\right]\Delta^{-1}, \tag{13}$$

and $\Delta = \mathbb{E}X_1 X_1^T$ *is the second moment matrix of* F*.*

Proof This theorem follows by writing, after adding and subtracting appropriate quantities,

$$\sqrt{n}(Q\hat{w}_{\mathrm{LS}} - \bar{w}) = \sqrt{n}S^{-1/2}U^T(\vec{a} - X\bar{w}) + \sqrt{n}\vec{h}_n,$$

where $\vec{h}_n \in \mathbb{R}^d$. Lemma 32 shows that the former of these terms converges in law to a normal. Using arguments similar to those in Theorem 7, we can show that $\sqrt{n}\vec{h}_n$ converges to zero in probability, and applying Slutsky's lemma completes the proof. Details can be found in Appendix E. ■

If the latent position $\bar{w}$ of the OOS vertex v is itself distributed according to F, integrating $\bar{w}$ above with respect to F yields the following corollary.

Corollary 12 *Assume the same setup as Theorem 11, but suppose that the true latent position of the out-of-sample vertex* v *is given by* $\bar{w} \sim F$*, independent of* (A, X)*. Then there exists a sequence of orthogonal matrices* $Q \in \mathbb{R}^{d\times d}$ *such that*

$$\sqrt{n}(Q\hat{w}_{\mathrm{LS}} - \bar{w}) \xrightarrow{\mathcal{L}} \int \mathcal{N}(0, \Sigma_{F,w})dF(w),$$

where $\Sigma_{F,w}$ *is as defined in Equation* (13)*. That is,* $\sqrt{n}(Q\hat{w}_{\mathrm{LS}} - \bar{w})$ *converges in distribution to a mixture of normals with mixing distribution* F*.*

Turning our attention to the LSE, we can obtain a similar CLT result for the LSE OOS extension, once we adjust for the fact that the LSE does not estimate the latent position $\bar{w}$ but instead estimates the vector $\tilde{w} = \bar{w}/\sqrt{n\mu^T\bar{w}}$, where $\mu \in \mathbb{R}^d$ is the mean of the inner-product distribution F. We note that the scaling of $\tilde{w}$ by the square root of the expected degree means that we must scale by n instead of the $\sqrt{n}$ scaling in the ASE CLTs above.

Theorem 13 *Let F be a d-dimensional inner-product distribution for which there exists a constant $\eta > 0$ such that $\eta \leq x^T y \leq 1-\eta$ whenever $x, y \in \operatorname{supp} F$. Let $(A, X) \sim \mathrm{RDPG}(F, n)$ and let v be the out-of-sample vertex with latent position $\bar{w} \in \mathbb{R}^d$ satisfying $\eta < \bar{w}^T x < 1-\eta$ for all $x \in \operatorname{supp} F$. Let $\check{w}_{\mathrm{LS}} \in \mathbb{R}^d$ denote the least-squares OOS extension of LSE as defined in Equation (10). Then there exists a sequence of orthogonal matrices $\tilde{Q} \in \mathbb{R}^{d\times d}$ such that*

$$n(\tilde{Q}\check{w}_{\mathrm{LS}} - \tilde{w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tilde{\Sigma}_{F,\bar{w}}),$$

where for any $w \in \mathbb{R}^d$ satisfying $\eta < \bar{w}^T x < 1-\eta$ for all $x \in \operatorname{supp} F$, we define

$$\tilde{\Sigma}_{F,\bar{w}} = \mathbb{E}\left[\frac{X_j^T\bar{w}(1-X_j^T\bar{w})}{\mu^T\bar{w}}\left(\frac{\tilde{\Delta}^{-1}X_j}{X_j^T\mu} - \frac{\bar{w}}{2\mu^T\bar{w}}\right)\left(\frac{\tilde{\Delta}^{-1}X_j}{X_j^T\mu} - \frac{\bar{w}}{2\mu^T\bar{w}}\right)^T\right], \tag{14}$$

with $\tilde{\Delta} = \mathbb{E}X_1X_1^T/\mu^T X_1$.

Proof The proof follows by a similar argument to the proof of Theorem 11, though it requires a more careful analysis to control convergence of the vertex degrees. By adding and subtracting appropriate quantities, we write

$$n(\tilde{Q}\check{w}_{\mathrm{LS}} - \tilde{w}) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{(a_j - X_j^T\bar{w})}{\sqrt{\mu^T\bar{w}}}\left(\frac{X_j}{X_j^T\mu} - \frac{\tilde{\Delta}\bar{w}}{2\mu^T\bar{w}}\right) + n\vec{h}_n,$$

where $\vec{h}_n \in \mathbb{R}^d$. The former of these two right-hand quantities is a sum of n independent mean-zero random variables, and hence converges to a normal with covariance $\tilde{\Sigma}_{F,\bar{w}}$. The remainder term $\vec{h}_n$ is controlled by concentration inequalities similar to those used in the proof of Theorem 7. Details are given in Appendix F. ■

4. Experiments

In this section, we briefly explore our results through simulations, and then turn our attention to investigating the performance of the out-of-sample extension on real-world data.

4.1 Simulation: speed of convergence

We first give a brief exploration of how quickly the asymptotic distribution in Theorem 11 becomes a good approximation. Toward this end, let us consider a simple mixture of point masses, $F = F_{\lambda,x_1,x_2} = \lambda\delta_{x_1} + (1-\lambda)\delta_{x_2}$, where $x_1, x_2 \in \mathbb{R}^2$ and $\lambda \in (0,1)$. This corresponds

to a two-block stochastic block model (Holland et al., 1983), in which the block probability matrix is given by

$$\begin{bmatrix} x_1^T x_1 & x_1^T x_2 \\ x_1^T x_2 & x_2^T x_2 \end{bmatrix}.$$

Corollary 12 implies that if all latent positions (including the OOS vertex) are drawn according to F, then the OOS estimate should be distributed as a mixture of normals centered at x_1 and x_2, with respective mixing coefficients λ and $1 - \lambda$.

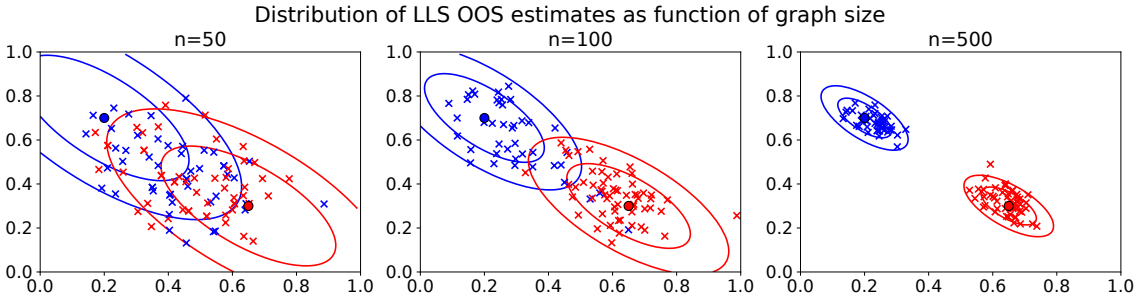


Figure 1: Observed distribution of the LLS OOS estimate for 100 independent trials for number of vertices $n = 50$ (left), $n = 100$ (middle) and $n = 500$ (right). Each plot shows the positions of 100 independent OOS embeddings, indicated by crosses, and colored according to cluster membership. Contours indicate two generalized standard deviations of the multivariate normal (i.e., 68% and 95% of the probability mass) about the true latent positions, which are indicated by solid circles. We note that even with merely 100 vertices, the normal approximation is already quite reasonable.

To assess how well the asymptotic distribution predicted by Theorem 11 and Corollary 12 holds, we generate RDPGs with latent positions drawn i.i.d. from distribution $F = F_{\lambda,x_1,x_2}$ defined above, with

$$\lambda = 0.4,\ x_1 = (0.2, 0.7)^T,\ \text{and}\ x_2 = (0.65, 0.3)^T.$$

For each trial, we draw $n + 1$ independent latent positions from F, and generate a binary adjacency matrix from these latent positions. We let the $(n+1)$-th vertex be the OOS vertex. Retaining the subgraph induced by the first n vertices, we obtain an estimate $\hat{X} \in \mathbb{R}^{n\times 2}$ via ASE, from which we obtain an estimate for the OOS vertex via the LS OOS extension as defined in (7). We remind the reader that for each RDPG draw, we initially recover the latent positions only up to a rotation. Thus, for each trial, we compute a Procrustes alignment (Gower and Dijksterhuis, 2004) of the in-sample estimates $\hat{X}$ to their true latent positions. This yields a rotation matrix R, which we apply to the OOS estimate. Thus, the OOS estimates are sensibly comparable across trials. Figure 1 shows the empirical distribution of the OOS embeddings of 100 independent RDPG draws, for $n = 50$ (left), $n = 100$ (center) and $n = 500$ (right) in-sample vertices. Each cross is the location of the OOS estimate for a single draw from the RDPG with latent position distribution F, colored

according to true latent position. OOS estimates with true latent position x_1 are plotted as blue crosses, while OOS estimates with true latent position x_2 are plotted as red crosses. The true latent positions x_1 and x_2 are plotted as solid circles, colored accordingly. The plot includes contours for the two normals centered at x_1 and x_2 predicted by Theorem 11 and Corollary 12, with the ellipses indicating the isoclines corresponding to one and two (generalized) standard deviations.

Examining Figure 1, we see that even with only 100 vertices, the mixture of normal distributions predicted by Theorem 11 holds quite well, with the exception of a few gross outliers from the blue cluster. With $n = 500$ vertices, the approximation is particularly good. Indeed, the $n = 500$ case appears to be slightly under-dispersed, possibly due to the Procrustes alignment. It is natural to wonder whether a similarly good fit is exhibited by the ML-based OOS extension. We conjectured at the end of Section 3 that a CLT similar to that in Theorem 11 would also hold for the ML-based OOS extension as defined in Equation (9). Figure 2 shows the empirical distribution of 100 independent OOS estimates, under the same experimental setup as Figure 1, but using the ML OOS extension rather than the linear least-squares extension. The plot supports our conjecture that the ML-based OOS estimates are also approximately normally distributed about the true latent positions. Broadly similar patterns hold for the same experiment applied to the least-squares LSE OOS extension, as predicted by Theorem 13.

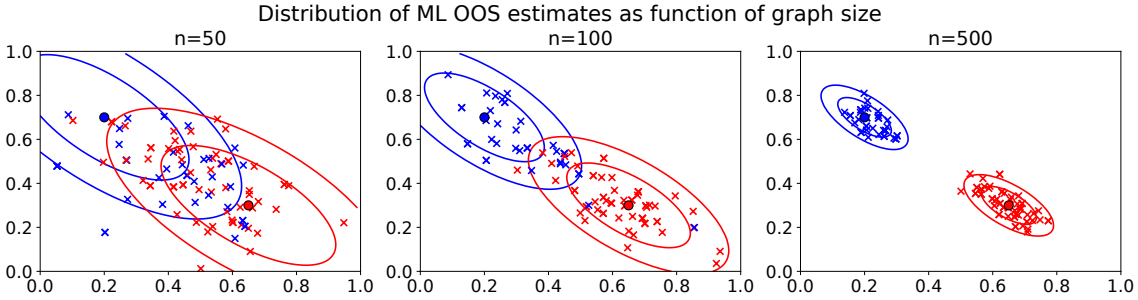


Figure 2: Observed distribution of the ML OOS estimate for 100 independent trials for number of vertices $n = 50$ (left), $n = 100$ (middle) and $n = 500$ (right). Each plot shows the positions of 100 independent OOS embeddings, indicated by crosses, and colored according to cluster membership. Contours indicate two generalized standard deviations of the multivariate normal about the true latent positions, which are indicated by solid circles. Once again, even with merely 100 vertices, the normal approximation is already quite reasonable, supporting our conjecture that the ML OOS estimates also distributed as a mixture of normals according to the latent position distribution F.

Figure 3 plots the same experiment as that performed in Figures 1 and 2, this time for the linear least squares OOS extension of the Laplacian spectral embedding. Recall that Theorem 13 predicts that the out-of-sample extension should be asymptotically normally distributed about the true (rescaled) latent position $\tilde{w} = \bar{w}/\sqrt{n\tilde{w}^T\mu}$. Compared to the previous two experiments, it is evident that the asymptotics are slightly slower to kick in,

but modulo the same Procrustes-induced underdispersion observed previously, the theorem appears to hold quite well with $n = 500$ vertices.

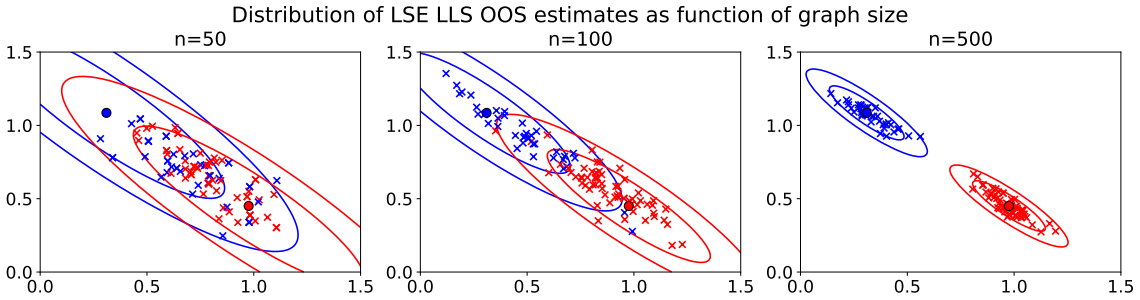


Figure 3: Observed distribution of the LSE OOS estimate for 100 independent trials for number of vertices $n = 50$ (left), $n = 100$ (middle) and $n = 500$ (right). Each plot shows the positions of 100 independent OOS embeddings, indicated by crosses, and colored according to cluster membership. Contours indicate two generalized standard deviations of the multivariate normal about the true latent positions, which are indicated by solid circles.

4.2 Tradeoff: computational cost versus classification accuracy

Figure 1 suggests that we may be confident in applying the large-sample approximation suggested by Theorem 11 and Corollary 12. Applying this approximation allows us to investigate the trade-offs between computational cost and classification accuracy, to which we now turn our attention. The mixture distribution F_{λ,x_1,x_2} above suggests a task in which, given an adjacency matrix A, we wish to classify the vertices according to which of two clusters or communities they belong. That is, we will view two vertices as belonging to the same community if their latent positions are the same (Holland et al., 1983, i.e., the latent positions specify an SBM,). More generally, one may view the task of recovering vertex block memberships in a stochastic block model as a clustering problem. Lyzinski et al. (2014) showed that applying ASE to such a graph, followed by k-means clustering of the estimated latent positions, correctly recovers community memberships of all the vertices (i.e., correctly assigns all vertices to their true latent positions) with high probability.

For concreteness, let us consider a still simpler mixture model, $F = F_{\lambda,p,q} = \lambda\delta_p + (1 - \lambda)\delta_q$, where $0 < p < q < 1$, and draw an RDPG $(\tilde{A}, X) \sim \text{RDPG}(F, n + m)$, taking the first n vertices to be in-sample, with induced adjacency matrix $A \in \mathbb{R}^{n\times n}$. That is, we draw the full matrix

$$\tilde{A} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where $C \in \mathbb{R}^{m\times m}$ is the adjacency matrix of the subgraph induced by the m OOS vertices and $B \in \mathbb{R}^{n\times m}$ encodes the edges between the in-sample vertices and the OOS vertices. The latent positions p and q encode a community structure in the graph $\tilde{A}$, and, as alluded to above, a common task in network statistics is to recover this community structure.

Let $\bar{w}^{(1)}, \bar{w}^{(2)}, \ldots, \bar{w}^{(m)} \in \{p, q\}$ denote the true latent positions of the m OOS vertices, with respective least-squares OOS estimates $\hat{w}^{(1)}_{\text{LS}}, \hat{w}^{(2)}_{\text{LS}}, \ldots, \hat{w}^{(m)}_{\text{LS}}$, each obtained from the in-sample ASE $\hat{X} \in \mathbb{R}^n$ of A. Corollary 12 implies that each $\hat{w}^{(t)}_{\text{LS}}$ for $t \in [m]$ is marginally (approximately) distributed as

$$\hat{w}^{(t)}_{\text{LS}} \sim \lambda \mathcal{N}(p, (n+1)^{-1}\sigma_p^2) + (1-\lambda)\mathcal{N}(q, (n+1)^{-1}\sigma_q^2),$$

where

$$\begin{aligned} \sigma_p^2 &= \Delta^{-2}\left(\lambda p^2(1-p^2)p^2 + (1-\lambda)pq(1-pq)q^2\right), \\ \sigma_q^2 &= \Delta^{-2}\left(\lambda pq(1-pq)p^2 + (1-\lambda)q^2(1-q^2)q^2\right), \\ \text{and } \Delta &= \lambda p^2 + (1-\lambda)q^2. \end{aligned}$$

Classifying the t-th OOS vertex based on $\hat{w}^{(t)}_{\text{LS}}$ according to the likelihood ratio thus has (approximate) probability of error

$$\eta_{n,p,q} = \lambda\left[1 - \Phi\left(\frac{\sqrt{n+1}(x_{n+1,p,q} - p)}{\sigma_p}\right)\right] + (1-\lambda)\Phi\left(\frac{\sqrt{n+1}(x_{n+1,p,q} - q)}{\sigma_q}\right),$$

where Φ denotes the cdf of the standard normal and $x_{n,p,q}$ is the value of x solving

$$\lambda\sigma_p^{-1}\exp\{n(x-p)^2/(2\sigma_p^2)\} = (1-\lambda)\sigma_q^{-1}\exp\{n(x-q)^2/(2\sigma_q^2)\},$$

and hence our overall error rate when classifying the m OOS vertices will grow as $m\eta_{n+1,p,q}$.

As discussed previously, the OOS extension allows us to avoid the expense of computing the ASE of the full matrix

$$\tilde{A} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}.$$

The LLS OOS extension is computationally inexpensive, requiring only the computation of the matrix-vector product $\hat{S}^{-1/2}\hat{U}^T\vec{a}$, with a time complexity $O(d^2n)$, assuming one does not precompute the product $\hat{S}^{-1/2}\hat{U}^T$. The eigenvalue computation required for embedding $\tilde{A}$ is thus far more expensive than the LLS OOS extension. Nonetheless, if one were intent on reducing the OOS classification error $\eta_{n+1,p,q}$, one might consider paying the computational expense of embedding $\tilde{A}$ to obtain estimates $\tilde{w}^{(1)}, \tilde{w}^{(2)}, \ldots, \tilde{w}^{(m)}$ of the m OOS vertices. That is, we obtain estimates for the m OOS vertices by making them in-sample vertices, at the expense of solving an eigenproblem on the $(m+n)$-by-$(m+n)$ adjacency matrix. Of course, the entire motivation of our approach is that the in-sample matrix A may not be available. Nonetheless, a comparison against this baseline, in which all data is used to compute our embeddings, is instructive.

Theorem 1 in Athreya et al. (2016) implies that the $\tilde{w}^{(t)}$ estimates, based on embedding the full matrix $\tilde{A}$, are (approximately) marginally distributed as

$$\tilde{w}^{(t)} \sim \lambda \mathcal{N}(p, (n+m)^{-1}\sigma_p^2) + (1-\lambda)\mathcal{N}(q, (n+m)^{-1}\sigma_q^2),$$

with classification error

$$\eta_{n+m,p,q} = \lambda\Phi\left(\frac{p - x_{n+m,p,q}}{\sigma_p}\right) + (1-\lambda)\Phi\left(\frac{x_{n+m,p,q} - q}{\sigma_q}\right),$$

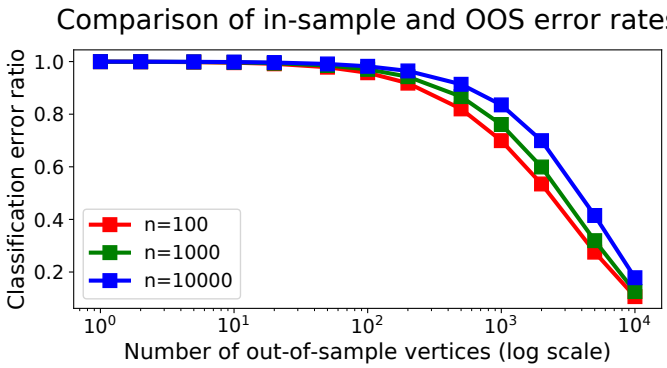


Figure 4: Ratio of the OOS classification error to the in-sample classification error as a function of the number of OOS vertices m, for $n = 100$ vertices, $n = 1,000$ vertices and $n = 10,000$ vertices. We see that for $m \leq 100$, the expensive in-sample embedding does not improve appreciably on the OOS classification error. However, when many hundreds or thousands of OOS vertices are available simultaneously (i.e., $m \geq 100$), we see that the in-sample embedding may improve upon the OOS estimate by a significant multiplicative factor.

where $x_{n+m,p,q}$ is the value of x solving

$$\lambda\sigma_p^{-1}\exp\{(m+n)(x-p)^2/(2\sigma_p^2)\} = (1-\lambda)\sigma_q^{-1}\exp\{(m+n)(x-q)^2/(2\sigma_q^2)\},$$

and it can be checked that $\eta_{n+m,q,p} < \eta_{n,q,p}$ when $m > 1$. Thus, at the cost of computing the ASE of $\tilde{A}$, we may obtain a better estimate. How much does this additional computation improve classification the OOS vertices? Figure 4 explores this question.

Figure 4 compares the error rates of the in-sample and OOS estimates as a function of m and n in the model just described, with $\lambda = 0.4, p = 0.6$ and $q = 0.61$. The plot depicts the ratio of the (approximate) in-sample classification error $\eta_{(n+m),p,q}$ to the (approximate) OOS classification error $\eta_{(n+1),p,q}$, as a function of the number of OOS vertices m, for differently-sized in-sample graphs, $n = 100; 1,000;$ and $10,000$. We see that over several magnitudes of graph size, the in-sample embedding does not improve appreciably over the OOS embedding except when multiple hundreds of OOS vertices are available. When hundreds or thousands of OOS vertices are available simultaneously, we see in the right-hand side of Figure 4 that the in-sample embedding classification error may improve upon the OOS classification error by a large multiplicative factor. Whether or not this improvement is worth the additional computational expense will depend upon the available resources and desired accuracy. For example, a researcher with access to specialized hardware (Zheng et al., 2015) or a researcher requiring a high degree of classification accuracy may be more willing to pay the computational expense. Nonetheless, Figure 4 suggests that the additional expense associated with performing a second ASE computation is only worthwhile in the event that hundreds or thousands of OOS vertices are available simultaneously. This surfeit of OOS vertices is rather different from the typical setting of OOS extension problems, where one typically wishes to embed at most a few previously unseen observations.

4.3 MNIST Digit Classification

We now consider, briefly, an application of our out-of-sample embedding to the MNIST data set (Lecun et al., 1998). This data set consists of 70,000 28-by-28 (i.e., 784-pixel) grey-scale images of hand-drawn digits, along with their digit labels (integers $0, 1, 2, \ldots, 9$), split into a training set of 55,000 images, a validation set of 5,000 images, and a test set of 10,000 images. Our goal is to predict the digit label given a grey-scale image.

To compare the out-of-sample extension to its in-sample counterpart, we consider the following set-up. We first choose a similarity measure κ given by

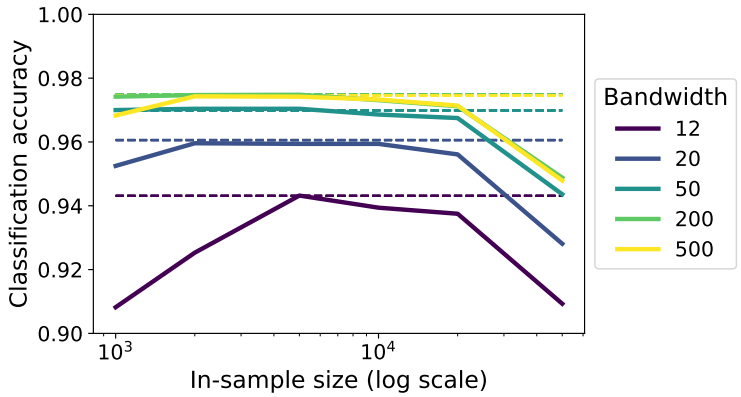
$$\kappa(i,j) = \exp\left\{\frac{-\|Z_i - Z_j\|^2}{\sigma}\right\}, \tag{15}$$

where $Z_i, Z_j \in \mathbb{R}^{784}$ are vectorized versions of images i and j, and $\sigma > 0$ is a bandwidth parameter. We construct a similarity matrix $K \in \mathbb{R}^{65,000\times 65,000}$ on all $65,000$ observations from the train and test set using the similarity measure κ as our kernel. Using a random sample of 1000 points from the validation data, we apply the elbow-finding technique of Zhu and Ghodsi (2006) to select an embedding dimension $\hat{d}$. The table on the right-hand side of Figure 5 summarizes the resulting embedding dimensions for different values of the bandwidth σ. Having chosen an embedding dimension, we embed the training and test data based on the similarity matrix K to obtain a gold standard in-sample embedding of the full 65,000-observation data set. Letting $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{55,000} \in \mathbb{R}^{\hat{d}}$ denote the resulting embeddings, we train a classifier based on the pairs $(\hat{X}_i, Y_i)_{i=1}^{55,000}$, where $Y_1, Y_2, \ldots, Y_{55,000} \in \{0, 1, 2, \ldots, 9\}$ are the labels of the training set images, and use it to predict the labels of the 10,000 test points. We refer to this as the *full classifier*. In what follows, we use a k-nearest neighbor classifier, with $k = 5$ in the experiments reported below, but performance was very similar for all choices of $k \in \{1, 3, 5, 7, 9, 11\}$.

For a given in-sample size $n < 55,000$ (in the experiments that follow, we in-sample sizes $n = 1,000,\ 2,000,\ 5,000,\ 10,000,\ 20,000,\ 50,000$), we select n observations from the training set uniformly at random without replacement and embed them into $\mathbb{R}^{\hat{d}}$ based on their induced submatrix of K (we use the ASE in the experiments reported below; LSE showed broadly similar behavior). Denote these points by $\mathring{X}_1, \mathring{X}_2, \ldots, \mathring{X}_n \in \mathbb{R}^{\hat{d}}$. We embed the remaining $m = 55,000 - n$ training examples according to the least squares ASE out-of-sample extension, and denote these points by $\check{X}_1, \check{X}_2, \ldots, \check{X}_m \in \mathbb{R}^{\hat{d}}$. We then train a classifier based on the embeddings $\{\mathring{X}_i\}_{i=1}^n \cup \{\check{X}_i\}_{i=1}^m$ and their corresponding digit labels. Embedding the $10,000$ observations in the test set via out-of-sample extension, we use the OOS classifier to predict their labels. We call this the *OOS classifier*.

If the full classifier and the OOS classifier obtain similar classification accuracy, we may conclude that the OOS extension has successfully captured the information present in the in-sample observations at a fraction of the computational cost. We note that our goal in this experiment is not necessarily to attain higher classification accuracy than existing state-of-the-art methods on the MNIST data set. Rather, we are using this well-studied data set to investigate the extent to which the out-of-sample extension introduced in this paper obtains performance comparable to the more expensive full embedding when applied to a real-world data set. That is, we are interested not in the absolute classification accuracy, but rather in the gap between the OOS classifier and the computationally costly full classifier.

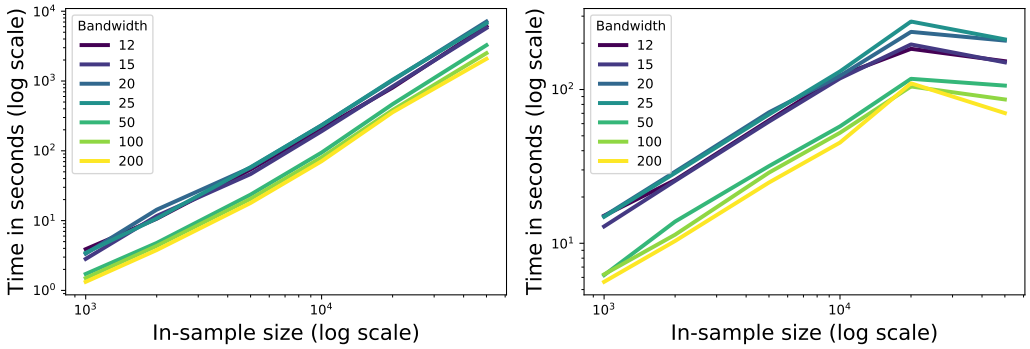
The plot on the left-hand side of Figure 5 compares the performance of the full and OOS classifiers for different values of the bandwidth σ. Each color corresponds to a choice of σ, with the solid lines indicating the classification accuracy of the OOS classifier as a function of the in-sample size. The horizontal dashed lines indicate the performance of the corresponding full classifier for each choice of bandwidth σ. Note that the dashed lines are all flat as a function of the in-sample size because the full classifier is trained on an embedding of all 65,000 observations (55,000 train and 10,000 test). Most importantly for our purposes, we see that for all choices of σ, there exists a choice of in-sample size for which the OOS classifier matches the performance of the full classifier, indicating that the computational savings of the OOS embedding do not necessarily come at the cost of downstream performance. Interestingly, it is not necessarily the case that a larger in-sample size yields better OOS classification accuracy.



σ	$\hat{d}$
12	95
15	89
20	105
25	107
50	44
100	38
200	33
300	33
500	33

Figure 5: Left: classification accuracy as a function of the in-sample size for different choices of bandwidth parameter σ. Performance of the OOS classifier is indicated by solid lines. Dashed lines indicate the performance of the full classifier. Right: embedding dimension $\hat{d}$ selected by the elbow-finding algorithm of Zhu and Ghodsi (2006) applied to 1,000 observations from the validation set, for different choices of bandwidth σ.

Broadly speaking, we see that smaller bandwidths generally yield poorer performance, likely due to the fact that when σ is small, the resulting similarity graph K is comparatively sparse, and more samples are required to adequately capture the geometry of the data. The fact that OOS classification accuracy decreases slightly for larger in-sample sizes points to a trade-off between the size of the in-sample and out-of-sample portions of the training set. A larger collection of in-sample training examples will tend to more accurately reflect the overall network topology. On the other hand, we have observed that the out-of-sample embeddings tend to be distributed slightly differently from the in-sample embeddings. We conjecture that this distributional difference is driven by the fact that the in-sample embeddings depend on the diagonal entries of K (recall that the contribution of these on-diagonal entries is asymptotically negligible), which we have taken here to be identically 1, to agree with Equation (15). In contrast, the OOS embeddings have no analogous parameter. The question of how best to choose the on-diagonal entries of the adjacency matrix in the ad-



(a) (b)

Figure 6: Computation time, in seconds, as a function of the in-sample size for different choice sof bandwidth σ, on a log-log scale. Each data point is the mean of ten independent experiment runs. (a) Time to compute the in-sample embedding. (b) Time to compute the out-of-sample extension.

jacency spectral embedding has been investigated extensively (see Section III.B in Tang et al., 2019, and citations therein), and it is not necessarily the case that 0 or 1 is the most appropriate choice for these entries. We leave for future work the question of how best to choose the on-diagonal entries of the in-sample adjacency matrix for subsequent out-of-sample embedding.

Figure 5 demonstrates that up to an in-sample size of approximately 10,000, the availability of more in-sample observations yields better classification accuracy across a range of bandwidth choices. However, larger in-sample sizes require greater computational resources due both to constructing the pairwise similarities among the in-sample set and to computing the leading eigenvectors of that matrix to construct the adjacency spectral embedding. These computational costs are demonstrated in Figure 6, which shows the computational time requirements for different in-sample set sizes. The left-hand plot shows the time required to construct the in-sample embeddings as a function of the in-sample size, while the right-hand plot shows the time required to construct the out-of-sample embeddings, also as a function of the in-sample size. Comparing the two plots, we see that constructing the in-sample embedding accounts for the vast majority of the computational time required by our experiment. This is as expected, in light of the fact that the out-of-sample extension avoids a costly eigenvalue problem. Most importantly, examining the left-hand plot in the context of Figure 5, we see that larger in-sample sizes require orders of magnitude more computation time with no discernible gain in classification accuracy, in keeping with our initial motivations for using the out-of-sample extension.

5. Discussion and Conclusion

We have presented theoretical results for out-of-sample extensions of graph embeddings, the adjacency spectral embedding and the Laplacian spectral embedding. In both cases, we have shown that under the random dot product graph, a least squares-based OOS

extension recovers the true latent position at the same rate as the more expensive in-sample embedding. Further, this linear least squares OOS extension obeys a CLT, whereby the OOS embedding is normally distributed about the true latent position. We have also presented results for an ASE OOS extension based on a maximum-likelihood objective function showing that this embedding recovers the true out-of-sample latent position at the same rate as the in-sample embedding. Experiments suggest that convergence to the predicted normal distribution is fairly fast, being a good approximation with only a few hundred vertices. Finally, we have briefly investigated how the approximation introduced by these OOS extensions might be traded off against the computational expense associated with computing the more expensive full graph embedding by investigating how the approximate classification error predicted by our CLT depends on the size of the in-sample graph and the number of out-of-sample vertices.

The results in this work suggest a number of interesting directions for future work, a few of which we briefly enumerate here. Firstly, though all of the OOS extensions presented in this paper match the asymptotic estimation error rates of their respective in-sample embeddings, our results say little about the constants associated with those rates or about finite-sample behavior of those OOS extensions (aside from their obvious restatements as finite-sample results alluded to briefly in Section 1.2). A more thorough investigation of how these different OOS extensions behave for different sizes of the in-sample graph and for different latent position distributions F would be of particular interest to practitioners faced with choosing between these different embeddings and OOS extensions as they apply to real data. Our discussion surrounding Figure 4 makes an initial step in this direction, but only suggests rules of thumb for when the speed/accuracy trade-off associated with out-of-sample extension is likely to be favorable.

A related line of questioning concerns how one should, when possible, select the in-sample vertices so as to yield optimal (as measured by, e.g., vertex classification or estimation of the latent positions) out-of-sample embeddings. Consider the setting where one has a graph $\tilde{G}$ of size $\tilde{n} = n + m$ that is far too large to be embedded via ASE or LSE. If n is the largest number of vertices that can be feasibly embedded as a full in-sample graph, it is natural to choose n vertices from $\tilde{G}$ to serve as the in-sample vertices, and embed the remaining m vertices via one of the out-of-sample extensions discussed in this paper. In this setting, how should one choose these n vertices from $\tilde{G}$? Problems of a similar nature have been considered elsewhere in the literature under the heading of *anchor graphs* or choosing *anchor points* (see, e.g., Liu et al., 2010), but we are not aware of any work in this area as it pertains to the ASE and LSE. This also suggests the problem of how best to embed m out-of-sample vertices jointly, rather than applying an OOS extension to each of them in isolation, particularly in the setting where we have access to the subgraph induced by these m out-of-sample vertices. Of most import here is the question, also explored by Figure 4, of how large the out-of-sample size m must be before one should prefer the expense of the full-graph embedding, and whether an embedding that makes use of this out-of-sample induced graph might bridge the gap between these two extremes by providing an embedding which, while more expensive than performing m OOS extensions in isolation, is still far less computationally intensive than embedding a graph of size $m + n$. A more thorough exploration of this trade-off from both a theoretical and empirical standpoint is the subject of on-going work.

Yet another line of inquiry concerns model misspecification. No network observed in nature is truly generated according to an RDPG, and it is natural to ask how the OOS extensions presented in this work fare in the presence of gross model misspecification. The RDPG (or its generalization, Rubin-Delanchy et al., 2017) can approximate any graphon, but it is less well-suited to capture, for example, scale-free properties exhibited by preferential attachment models (Barabási and Albert, 1999; Dorogovtsev et al., 2000). In preliminary experiments, we have found that even under severe model misspecification (e.g., applying an RDPG to data generated from a preferential attachment model), the in-sample and out-of-sample embeddings are quite similar. That is, we have weak evidence that results similar to our concentration results presented in Section 3.1 hold even when the RDPG assumption is violated. Unfortunately, to the best of our knowledge, there do not exist any scale-free network models in which there is also a natural notion of latent geometry or cluster membership (a possible exception is the hyperbolic geometric random graph, Krioukov et al., 2010, but the non-Euclidean geometry of the latent positions makes estimation accuracy difficult to assess). As a result, it is hard to say how large or small an error between the in-sample and out-of-sample embeddings is operationally significant, rendering these experiments rather uninformative. Further, the similarity of the in-sample and out-of-sample embeddings need not depend in any way on how well the network itself is captured by the RDPG's low-rank assumption. We leave for future work a more thorough exploration of the consequences of model misspecification for the OOS problem, and for inference under the RDPG more generally.

Finally, we note the connections between the present work and other large-scale dimensionality reduction problems, particularly distributed algorithms for matrix factorization (Ahmed et al., 2013; Fan et al., 2019). The data access model we have considered is rather different from that assumed in most work on distributed computing, but future work will explore distributed versions of the OOS extensions developed here, starting by applying known methods for distributed least squares estimation (Xiao et al., 2005).

Acknowledgments

KL was supported by NSF grants DMS-1646108 and DMS-2052632, as well as by the University of Wisconsin-Madison OVCRGE with funding from the Wisconsin Alumni Research Foundation. FR is grateful for the support by the Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) as well as Discovery Early Career Researcher Award (DE180100923). MT and CEP were supported by DARPA grant FA8750-17-2-0112. MWM acknowledges ARO, DARPA, NSF, and ONR for providing partial support of this work.

Appendix A. Technical Results for the Random Dot Product Graph

Here we collect a number of basic results that will be useful in our subsequent proofs of the main theorems. Most of the results in this section are adapted from existing results in Levin et al. (2017), Lyzinski et al. (2014) and Tang and Priebe (2018). We refer the

interested reader to Athreya et al. (2018) for a more thorough overview of the RDPG and the statistical problems that arise in relation to it.

Lemma 14 (Levin et al. (2017), Observation 2) *Let $(A, X) \sim \mathrm{RDPG}(F, n)$ for some d-dimensional inner product distribution F. There exists constants $0 < C_1 < C_2$, depending only on F, such that with probability 1 it holds for all suitably large n that*

$$C_1 n \leq \lambda_d(P) \leq \lambda_1(P) \leq C_2 n \text{ and}$$
$$C_1\sqrt{n} \leq \lambda_d(X) \leq \lambda_1(X) \leq C_2\sqrt{n}.$$

Lemma 15 (Levin et al. (2017), Lemma 3) *With notation as above, let $V_1 \Lambda V_2^T$ be the SVD of $U^T\hat{U} \in \mathbb{R}^{d\times d}$, and define $Q = V_1 V_2^T$. Then*

$$\|U^T\hat{U} - Q\|_F = O(n^{-1}\log n).$$

Lemma 16 (Tang and Priebe (2018), Proposition B.2) *With notation as above, let $\tilde{V}_1\tilde{\Lambda}\tilde{V}_2^T$ be the SVD of $\tilde{U}^T\check{U} \in \mathbb{R}^{d\times d}$ and define $\tilde{Q} = \tilde{V}_1\tilde{V}_2^T$. Then*

$$\|\tilde{U}^T\check{U} - \tilde{Q}\|_F = O(n^{-1}).$$

Lemma 17 (Lyzinski et al. (2017) Lemma 15; Tang and Priebe (2018) Lemma B.3) *With notation as above,*

$$\left\|\check{U}^T\tilde{U}\tilde{S}^{1/2} - \check{S}^{1/2}\check{U}^T\tilde{U}\right\| = O(n^{-1}),$$
$$\left\|\hat{U}^T U S^{-1/2} - \hat{S}^{-1/2}\hat{U}^T U\right\|_F = O(n^{-3/2}\log n) \text{ and}$$
$$\left\|\hat{U}^T U S^{1/2} - \hat{S}^{1/2}\hat{U}^T U\right\|_F = O(n^{-1/2}\log n).$$

Lemma 18 *Let F be a d-dimensional inner-product distribution and let $(A, X) \sim \mathrm{RDPG}(F, n)$, and let v be the out-of-sample vertex with latent position $\bar{w} \in \mathbb{R}^d$ such that $0 \leq \bar{w}^T x \leq 1$ for all $x \in \operatorname{supp} F$. For $i \in [n]$, let $d_i = \sum_j A_{i,j}$ denote the degree of vertex i and $t_i = \sum_j X_j^T X_i = \mathbb{E}[d_i|X]$ denote its expectation conditional on the latent positions. Analogously, let $d_v = \sum_j a_j$ denote the degree of the out-of-sample vertex and $t_v = \sum_j X_j^T\bar{w}$ denote its expectation. Then*

$$\max\{|d_i - t_i| : i \in [n] \cup \{v\}\} = O(\sqrt{n}\log^{1/2} n). \tag{16}$$

Similarly, letting $\mu = \mathbb{E}X_1 \in \mathbb{R}^d$ denote the mean of latent position distribution F and taking $X_v = \bar{w}$,

$$\max\left\{\left|t_i - n\mu^T X_i\right| : i \in [n] \cup \{v\}\right\} = O(\sqrt{n}\log^{1/2} n). \tag{17}$$

Further, uniformly over all $i \in [n]$,

$$|d_i^{-1/2} - t_i^{-1/2}| = O(n^{-1}\log^{1/2} n), \tag{18}$$
$$|d_i^{-1} - t_i^{-1}| = O(n^{-3/2}\log^{1/2} n), \tag{19}$$
$$t_i = \Theta(n) \tag{20}$$

Proof Fix some $i \in [n] \cup \{v\}$. By definition, we have

$$d_i - t_i = \begin{cases} \sum_{j \neq i}(A_{i,j} - P_{i,j}) & \text{if } i \in [n] \\ \sum_{j=1}^{n} a_j - X_j^T \bar{w} & \text{if } i = v, \end{cases}$$

a sum of independent random variables, each contained in $[-1, 1]$ and thus Hoeffding's inequality immediately yields

$$\Pr[|d_i - t_i| \geq s] \leq 2 \exp\left\{\frac{-2s^2}{n}\right\}$$

for any $s \geq 0$. Taking $s = C\sqrt{n}\log^{1/2} n$ for suitably large constant $C > 0$, we have

$$\Pr\left[|d_i - t_i| \geq C\sqrt{n}\log^{1/2} n\right] \leq C' n^{-3}.$$

Taking a union bound over all $i \in [n] \cup \{v\}$, we conclude that

$$\Pr\left[\exists i : |d_i - t_i| \geq C\sqrt{n}\log^{1/2} n\right] \leq C n^{-2},$$

and an application of the Borel-Cantelli Lemma (Billingsley, 1995) yields Equation (16).

Again by definition, we have for any $i \in [n] \cup \{v\}$,

$$t_i - n X_i^T \mu = X_i^T (X_i - \mu) + \sum_{j \neq i} X_i^T (X_j - \mu).$$

The first term on the right-hand side is $O(1)$, since $X_i \sim F$ and μ is constant. The sum over $j \neq i$ is, conditioned on X_i, a sum of independent unbiased random variables, which are bounded by the assumption that $0 \leq x^T y \leq 1$ whenever $x, y \in \operatorname{supp} F$. Thus, an application of Hoeffding's inequality similar to that above yields that, conditioned on $X_i = x_i \in \operatorname{supp} F$,

$$\sum_{j \neq i} x_i^T (X_j - \mu) \leq C\sqrt{n}\log^{1/2} n,$$

where the constant C can be chosen independent of x_i again because $\operatorname{supp} F$ is bounded. Unconditioning establishes Equation (17), since $X_i^T(X_i - \mu) = O(1)$. (20) follows, since $t_i = n X_i^T \mu + O(\sqrt{n}\log^{1/2} n)$. Writing

$$\left|\frac{1}{\sqrt{d_i}} - \frac{1}{\sqrt{t_i}}\right| = \frac{|d_i - t_i|}{\sqrt{d_i}\sqrt{t_i}(\sqrt{d_i} + \sqrt{t_i})}$$

and applying Equations (17) and (20) implies (18). A similar argument establishes (19). ■

Lemma 19 *Let $P = XX^T \in \mathbb{R}^{n \times n}$ with rows of X drawn i.i.d. from F as above. Then*

$$\lambda_d(\mathcal{L}(P)) = \Theta(1), \lambda_1(\mathcal{L}(P)) = \Theta(1) \text{ and } \lambda_d(\tilde{X}) = \Theta(1). \tag{21}$$

Proof By definition, $\mathcal{L}(P) = T^{-1/2}USU^TT^{-1/2}$, so that

$$\lambda_d(\mathcal{L}(P)) \leq \lambda_1(\mathcal{L}(P)) \leq \|\mathcal{L}(P)\| \leq \|T^{-1/2}\|\|S\|\|T^{-1/2}\| \leq \frac{\|S\|}{\min_i t_i} \leq C,$$

where the last inequality follows from Lemmas 14 and 18.

To show the corresponding lower-bound, we adapt an argument from the proof of Theorem 8.1.17 in Golub and Van Loan (2012) to write

$$\lambda_1^2(T^{1/2})\lambda_d(\mathcal{L}(P)) \geq \lambda_d(P) \geq Cn,$$

where the second lower-bound follows from Lemma 14. We conclude that

$$\lambda_1(\mathcal{L}(P)) \geq \lambda_d(\mathcal{L}(P)) \geq \frac{Cn}{\lambda_1(T)} \geq C,$$

since $\lambda_1^2(T^{1/2}) = \lambda_1(T) \leq n$.

By definition of $\tilde{X}$, $\lambda_k(\tilde{X}) = \sqrt{\lambda_k(\mathcal{L}(P))}$ for all $k \in [d]$, whence $\lambda_d(\tilde{X}) = \Theta(1)$ ■

Lemma 20 *Let F be a d-dimensional inner-product distribution with mean μ and suppose that there exists a constant $\eta > 0$ such that $\eta \leq x^Ty \leq 1 - \eta$ for all $x, y \in \operatorname{supp} F$. Define $\tilde{\Delta} = \mathbb{E}X_1X_1^T/X_1^T\mu$ where $X_1 \sim F$ and let $\check{S} = \check{X}^T\check{X}$ and $\tilde{S} = \tilde{X}^T\tilde{X}$. Then*

$$\|\tilde{Q}\check{S}\tilde{Q}^T - \tilde{S}\| = O\left(\frac{1}{n}\right) \text{ and } \|\tilde{S} - \tilde{\Delta}\| = O\left(\frac{\log^{1/2} n}{\sqrt{n}}\right).$$

Proof Adding and subtracting appropriate quantities and applying a triangle inequality followed by submultiplicativity, we have

$$\begin{aligned}\|\tilde{Q}\check{S}\tilde{Q}^T - \tilde{S}\| &= \left\|\tilde{Q}\check{S}^{1/2}\left(\check{S}^{1/2}\tilde{Q}^T - \tilde{Q}^T\tilde{S}^{1/2}\right) + \left(\tilde{Q}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{Q}\right)\tilde{Q}^T\tilde{S}^{1/2}\right\| \\ &\leq \left(\|\tilde{Q}\check{S}^{1/2}\| + \|\tilde{Q}^T\tilde{S}^{1/2}\|\right)\|\tilde{Q}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{Q}\|,\end{aligned}$$

where we have used the unitary invariance of the spectral norm to write

$$\|\tilde{Q}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{Q}\| = \|\check{S}^{1/2}\tilde{Q}^T - \tilde{Q}^T\tilde{S}^{1/2}\|.$$

An additional application of the unitary invariance of the spectral norm yields

$$\|\tilde{Q}\check{S}\tilde{Q}^T - \tilde{S}\| \leq \left(\|\check{S}^{1/2}\| + \|\tilde{S}^{1/2}\|\right)\|\tilde{Q}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{Q}\|. \tag{22}$$

By definition of $\check{S}$ and $\tilde{S}$ as the top d eigenvalues of $\mathcal{L}(A)$ and $\mathcal{L}(P)$, respectively, we have

$$\|\check{S} - \tilde{S}\| \leq \|\mathcal{L}(A) - \mathcal{L}(P)\|.$$

Theorem 3.1 in Oliveira (2010) implies that

$$\|\mathcal{L}(A) - \mathcal{L}(P)\| \leq C\left(\min_i t_i\right)^{-1/2}\log^{1/2} n,$$

and Lemma 18 implies that $\min_i t_i = \Omega(n)$, so that

$$\|\mathcal{L}(A) - \mathcal{L}(P)\| = O(n^{-1/2}\log^{1/2} n),$$

and it follows that

$$\|\check{S}^{1/2}\| \leq= \|\tilde{S}^{1/2}\|\,(1 + o(1))\,.$$

Lemma 19 bounds the growth of $\|\tilde{S}\|$ as $O(1)$, whence $\|\check{S}^{1/2}\| = O(1)$ and we conclude that

$$\|\check{S}^{1/2}\| + \|\tilde{S}^{1/2}\| = O(1). \tag{23}$$

Once again adding and subtracting appropriate quantities, applying the triangle inequality folowed by submultiplicativity,

$$\begin{aligned}\|\tilde{Q}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{Q}\| &\leq \|(\tilde{Q} - \tilde{U}^T\check{U})\check{S}^{1/2}\| + \|\tilde{U}^T\check{U}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{U}^T\check{U}\| + \|\tilde{S}^{1/2}(\tilde{U}^T\check{U} - \tilde{Q})\| \\ &\leq \left(\|\check{S}^{1/2}\| + \|\tilde{S}^{1/2}\|\right)\|\tilde{Q} - \tilde{U}^T\check{U}\| + \|\tilde{U}^T\check{U}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{U}^T\check{U}\|.\end{aligned}$$

Equation (23) and Lemma 16 imply that

$$\left(\|\check{S}^{1/2}\| + \|\tilde{S}^{1/2}\|\right)\|\tilde{Q} - \tilde{U}^T\check{U}\| = O(n^{-1}),$$

and Lemma 17 implies that

$$\|\tilde{U}^T\check{U}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{U}^T\check{U}\| = O(n^{-1}).$$

Combining the above two displays, we conclude that

$$\|\tilde{Q}\check{S}^{1/2} - \tilde{S}^{1/2}\tilde{Q}\| = O(n^{-1}).$$

Applying this and Equation (23) to Equation (22), we conclude that $\|\tilde{Q}\check{S}\tilde{Q}^T - \tilde{S}\| = O(n^{-1})$.

To bound $\|\tilde{S} - \tilde{\Delta}\|$, note that

$$\tilde{S} = \sum_{i=1}^{n} \tilde{X}_i\tilde{X}_i^T = \sum_{i=1}^{n} \frac{X_iX_i^T}{t_i}.$$

Applying Lemma 18, $\max_i |t_i^{-1} - (nX_i^T\mu)^{-1}| = O(n^{-3/2}\log^{1/2} n)$, and thus

$$\tilde{S} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_iX_i^T}{X_i^T\mu} + O(n^{-1/2}\log^{1/2} n).$$

Hoeffding's inequality applied to the sum, using our assumption that $\eta \leq X_iX_i^T/X_i^T\mu \leq 1 - \eta$ with probability 1, implies that $\tilde{S} = \tilde{\Delta} + O(n^{-1/2}\log^{1/2} n)$, completing the proof. ■

Lemma 21 *Suppose that F is a d-dimensional inner-product distribution with $X_1 \sim F$ for which $\Delta = \mathbb{E}_F X_1X_1^T \in \mathbb{R}^{d\times d}$ is full rank. If $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F$, then with probability 1 there exists an n_0 such that $X \in \mathbb{R}^{n\times d}$ has full column rank for all $n \geq n_0$.*

Proof Since the top d eigenvalues of $P = XX^T$ are precisely the d eigenvalues of X^TX, Lemma 14 implies that $\lambda_d(X^TX) = \Omega(n)$. It follows that $X^TX \in \mathbb{R}^{d\times d}$ is invertible for all suitably large n. ■

We now give a proof of the bound in Equation 6 in Lemma 6.

Proof [Proof of Lemma 6] Let $\zeta_i \in \mathbb{R}^d$ denote the (transposed) i-th row of $\check{X} - \tilde{X}\tilde{Q}$, where $\tilde{Q} = \tilde{V}_1\tilde{V}_2^T$ as in Lemma 16 above. Define the event

$$E_n = \left\{ \forall i \in [n] : \|\zeta_i\| \leq \frac{C \log^{1/2} n}{n} \right\}$$

where $C > 0$ is a constant that we will specify below, depending on the latent position distribution F but not on n. It will suffice for us to show that E_n holds eventually.

Fix some $i \in [n]$ and define $\mu = \mathbb{E}X_1 \in \mathbb{R}^d$ to be the mean of F. Following the argument in Appendix B.1 of Tang and Priebe (2018), we have

$$\zeta_i = \frac{(\tilde{X}^T\tilde{X})^{-1}}{n} \frac{\sqrt{n}}{\sqrt{t_i}} \sum_{j\neq i} \frac{A_{i,j} - P_{i,j}}{\sqrt{n}} \left(\frac{X_j}{X_j^T\mu} - \frac{\tilde{\Delta}X_i}{2X_i^T\mu} \right) + o(n^{-1}). \tag{24}$$

For all $j \in [n] \setminus \{i\}$, define

$$Z_j^{(i)} = \frac{A_{i,j} - P_{i,j}}{\sqrt{n}} \left(\frac{X_j}{X_j^T\mu} - \frac{\tilde{\Delta}X_i}{2X_i^T\mu} \right).$$

Condition on $X_i = x_i \in \operatorname{supp} F$ and fix $k \in [d]$. Thanks to the assumption that $0 < \eta \leq x^Ty \leq 1-\eta$ whenever $x, y \in \operatorname{supp} F$, we have that $\sum_{j\neq i} Z_{j,k}^{(i)}$ is a sum of independent mean-0 bounded random variables. Hoeffding's inequality implies that

$$\Pr\left[\left| \sum_{j\neq i} Z_{j,k}^{(i)} \right| \geq s \middle| X_i = x_i \right] \leq 2\exp\left\{ \frac{-s^2}{2n^{-1}\sum_{j\neq i} V_{j,k}^2} \right\}, \tag{25}$$

where

$$V_{j,k} = \frac{X_{j,k}}{X_j^T\mu} - \frac{(\tilde{\Delta}x_i)_k}{2x_i^T\mu}.$$

Using the fact that $X_j, x_i \in \operatorname{supp} F$ and that X_j is independent of X_i for $j \neq i$, we have

$$\mathbb{E}\left[V_{j,k}^2 \middle| X_i = x_i \right] \leq C \left(\frac{\|\tilde{\Delta}x_i\|^2}{4(x_i^T\mu)^2} + \mathbb{E}\left[\frac{|X_{j,k}|^2}{(X_j^T\mu)^2} \right] \right)^2 \leq C_F, \tag{26}$$

where C_F depends on F but can be chosen independent of k and x_i. By the law of large numbers (conditional on $X_i = x_i$),

$$n^{-1}\sum_{j\neq i} V_{j,k}^2 \to \mathbb{E}[V_{j,k}^2 \mid X_i = x_i] \text{ almost surely.}$$

Thus, applying Equation (26) and integrating out by X_i,

$$n^{-1}\sum_{j\neq i} V_{j,k}^2 \leq 2C_F \text{ eventually.}$$

Integrating (25) with respect to F and using the above fact, we conclude that

$$\Pr\left[\left|\sum_{j\neq i} Z_{j,k}^{(i)}\right| \geq C\log^{1/2} n\right] \leq 2n^{-3},$$

for suitably large constant $C > 0$. A union bound over all $k \in [d]$ yields

$$\Pr\left[\left\|\sum_{j\neq i} Z_{j}^{(i)}\right\| \geq C\log^{1/2} n\right] \leq 2dn^{-3},$$

and a further union bound over $i \in [n]$ implies

$$\max_{i\in[n]} \|\sum_{j\neq i} Z_j^{(i)}\| = O(\log^{1/2} n). \tag{27}$$

Applying this result to Equation (24) and using the fact that $\tilde{X}^T\tilde{X} \to \tilde{\Delta}$ almost surely and $\sqrt{n}t_i^{-1/2} = O(1)$ by Lemmas 18 and 20 respectively, we have

$$\max_{i\in[n]} \|\zeta_i\| \leq \frac{1}{n\|\tilde{X}^T\tilde{X}\|}\frac{\sqrt{n}}{\min_{i\in[n]}\sqrt{t_i}} \max_{i\in[n]}\left\|\sum_{j\neq i} Z_j^{(i)}\right\| = O\left(\frac{\log^{1/2} n}{n}\right), \tag{28}$$

which completes the proof. ∎

The following spectral norm bound will be useful at several points in our proofs.

Theorem 22 (Matrix Bernstein inequality, Tropp, 2015) *Let $\{Z_k\}$ be a finite collection of random matrices in $\mathbb{R}^{d_1\times d_2}$ with $\mathbb{E}Z_k = 0$ and $\|Z_k\| \leq R$ for all k, then*

$$\Pr\left[\left\|\sum_k Z_k\right\| \geq t\right] \leq (d_1+d_2)\exp\left\{\frac{-t^2}{\nu^2 + Rt/3}\right\},$$

where

$$\nu^2 = \max\left\{\left\|\sum_k \mathbb{E}Z_k Z_k^T\right\|, \left\|\sum_k \mathbb{E}Z_k^T Z_k\right\|\right\}.$$

Appendix B. Proof of ASE LS-OOS Concentration Inequality

To prove Theorem 7, we must relate the least squares solution $\hat{w}_{\text{LS}}$ of (7) to the true latent position $\bar{w}$. We will proceed in two steps. First, we will show that $\hat{w}_{\text{LS}}$ is close to a least

squares solution based on the true latent positions $\{X_i\}_{i=1}^n$ rather than on the estimates $\{\hat{X}_i\}_{i=1}^n$. That is, letting w_{LS} be the solution

$$w_{\text{LS}} = \arg\min_{w \in \mathbb{R}^d} \|Xw - \vec{a}\|_F, \tag{29}$$

we will bound the error introduced by the ASE, $\|Q\hat{w}_{\text{LS}} - w_{\text{LS}}\|$, taking $Q \in \mathbb{R}^{d\times d}$ to be as defined in Lemma 6. This is the content of Lemma 26. Second, we will show that w_{LS} is close to the true latent position $\bar{w}$. That is, we will control the error introduced by the n random in-sample latent positions and the network A. This is done in Lemma 27. The triangle inequality will then yield Theorem 7.

We first establish a bound on $\|Q\hat{w}_{\text{LS}} - w_{\text{LS}}\|$, where $\hat{w}_{\text{LS}}$ is the solution to Equation (7), w_{LS} is as defined by Equation (29), and $Q \in \mathbb{R}^{d\times d}$ is the orthogonal matrix guaranteed to exist by Lemma 6. Our bound will depend upon a basic result for solutions of perturbed linear systems, which we adapt from Golub and Van Loan (2012). In essence, we wish to compare

$$\hat{w}_{\text{LS}} = \arg\min_{w \in \mathbb{R}^d} \|\hat{X}w - \vec{a}\|_F$$

against

$$w_{\text{LS}} = \arg\min_{w \in \mathbb{R}^d} \|Xw - \vec{a}\|_F.$$

Recall that for a matrix $B \in \mathbb{R}^{n\times d}$ of full column rank, we define the condition number

$$\kappa_2(B) = \frac{\sigma_1(B)}{\sigma_d(B)}.$$

Theorem 23 (Golub and Van Loan (2012), Theorem 5.3.1) *Suppose that the quantities $w_{\text{LS}}, \hat{w}_{\text{LS}} \in \mathbb{R}^d$ and $r_{\text{LS}}, \hat{r}_{\text{LS}} \in \mathbb{R}^n$ satisfy*

$$\|Xw_{\text{LS}} - \vec{a}\| = \min_w \|Xw - \vec{a}\|, \quad r_{\text{LS}} = \vec{a} - Xw_{\text{LS}},$$
$$\|\hat{X}\hat{w}_{\text{LS}} - \vec{a}\| = \min_w \|\hat{X}w - \vec{a}\|, \quad \hat{r}_{\text{LS}} = \vec{a} - \hat{X}\hat{w}_{\text{LS}},$$

and that

$$\|\hat{X} - XQ\| < \lambda_d(X). \tag{30}$$

Assume $\vec{a}, r_{\text{LS}}$ and w_{LS} are all non-zero and define $\theta_{\text{LS}} \in (0, \pi/2)$ by $\sin\theta_{\text{LS}} = \|r_{\text{LS}}\|/\|\vec{a}\|$. Letting

$$\nu_{\text{LS}} = \frac{\|Xw_{\text{LS}}\|}{\sigma_d(XQ)\|Q^T w_{\text{LS}}\|},$$

we have

$$\begin{aligned}&\frac{\|\hat{w}_{\text{LS}} - Q^T w_{\text{LS}}\|}{\|Q^T w_{\text{LS}}\|} \\ &\leq \frac{\|\hat{X} - XQ\|}{\|XQ\|}\left(\frac{\nu_{\text{LS}}}{\cos\theta_{\text{LS}}} + (1 + \nu_{\text{LS}}\tan\theta_{\text{LS}})\kappa_2(XQ)\right) + O\left(\frac{\|\hat{X} - XQ\|^2}{\|XQ\|^2}\right).\end{aligned} \tag{31}$$

To apply Theorem 23, we will first need to show that the condition in Equation (30) and the non-zero conditions on $\vec{a}, r_{\text{LS}}$ and w_{LS} all hold with high probability. This is done in Lemma 24. We will then show, using Lemma 24 and Lemma 25, that the right-hand side of Equation (31) is $O(n^{-1/2}\log n)$.

Lemma 24 *With notation as above, $\vec{a}$, r_{LS} and w_{LS} are all nonzero eventually, and (30) holds eventually. That is, with probability 1, there exists a sequence of orthogonal matrices $Q \in \mathbb{R}^{d\times d}$ such that*

$$\|\hat{X} - XQ\| < \lambda_d(X) \text{ eventually.} \tag{32}$$

Further,

$$\frac{\|\hat{X} - XQ\|}{\|XQ\|} = O\left(\frac{\log n}{\sqrt{n}}\right). \tag{33}$$

Proof That $\vec{a}$ is non-zero eventually is an immediate consequence of the model, and it follows that w_{LS} is non-zero eventually, from which it follows that the residual $r_{\text{LS}} = \vec{a} - Xw_{\text{LS}}$ is also nonzero eventually. Let $Q \in \mathbb{R}^{d\times d}$ be the orthogonal matrix guaranteed by Lemma 6. We begin by observing that

$$\|\hat{X} - XQ\|^2 \leq \|\hat{X} - XQ\|_F^2 = \sum_{i=1}^{n} \|\hat{X}_i - QX_i\|^2 = O(\log^2 n)$$

where the last equality follows from Lemma 6. By the definition of the RDPG, we can write $XQ = US^{1/2}Q$, from which $\sigma_d(XQ) = \sigma_d^{1/2}(P) = \Omega(\sqrt{n})$ by Lemma 14. This establishes (32) immediately, and (33) follows from the above display. ■

Lemma 25 *With notation as in Theorem 23, there exists a constant $0 \leq \gamma < 1$, not depending on n, such that with probability 1, $\cos\theta_{\text{LS}} \geq \gamma$ for all suitably large n. That is, there exists a constant $0 < \gamma'$ such that*

$$\frac{\|XQw_{\text{LS}} - \vec{a}\|}{\|\vec{a}\|} \leq \gamma' \text{ eventually.}$$

Proof By definition of w_{LS}, we have $\|XQw_{\text{LS}} - \vec{a}\| \leq \|X\bar{w} - \vec{a}\|$. For ease of notation, set $\vec{r} = \vec{a} - X\bar{w}$. It will suffice for us to show that for some constant $\rho > 0$, we have

$$(1-\rho)\|\vec{a}\|^2 - \|\vec{r}\|^2 \geq 0 \text{ eventually,} \tag{34}$$

since then, after rearranging terms, $\sin^2\theta_{\text{LS}} \leq 1 - \rho$. To show (34), note that

$$\begin{aligned}(1-\rho)\|\vec{a}\|^2 - \|\vec{r}\|^2 &= 2\sum_{i=1}^{n} a_i X_i^T\bar{w} - \sum_{i=1}^{n}(X_i^T\bar{w})^2 - \rho\sum_{i=1}^{n} a_i^2 \\ &\geq \mathbb{E}\left[(1-\rho)\|\vec{a}\|^2 - \|\vec{r}\|^2\right] + C\sqrt{n}\log^{1/2} n \text{ eventually,}\end{aligned}$$

where the inequality follows from an application of Hoeffding's inequality to show that the sum concentrates about its expectation. We will have established (34) if we can show that

$\mathbb{E}\left[(1-\rho)\|\vec{a}\|^2 - \|\vec{r}\|^2\right]$ grows faster than $C\sqrt{n}\log^{1/2} n$. To establish this, let $i \in [n]$ be arbitrary and write

$$\begin{aligned}\mathbb{E}\left[(1-\rho)a_i^2 - r_i^2\right] &= \mathbb{E}\left[(1-\rho)a_i^2 - (a_i - X_i^T\bar{w})^2\right] = -\rho\mathbb{E}a_i^2 + 2\mathbb{E}a_i X_i^T\bar{w} - \mathbb{E}(X_i^T\bar{w})^2 \\ &= -\rho\mathbb{E}a_i^2 - \mathbb{E}(a_i - X_i^T\bar{w})X_i^T\bar{w} + \mathbb{E}a_i X_i^T\bar{w} = \mathbb{E}a_i X_i^T\bar{w} - \rho\mathbb{E}a_i^2.\end{aligned}$$

By our boundedness assumption on $\bar{w}$ and $\operatorname{supp} F$, $\mathbb{E}a_i X_i^T\bar{w} = \mathbb{E}(X_i^T\bar{w})^2$ is bounded away from zero uniformly in $i \in [n]$. Thus, choosing $\rho > 0$ suitably small ensures that there exists a small constant $\eta' > 0$ such that $\mathbb{E}\left[(1-\rho)a_i^2 - r_i^2\right] \geq \eta' > 0$. Summing over n,

$$\mathbb{E}\left[(1-\rho)\|\vec{a}\|^2 - \|\vec{r}\|^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(1-\rho)a_i^2 r_i^2\right] \geq n\eta' = \Omega(n),$$

which proves the bound in (34), completing the proof. ■

Lemma 26 *With notation as in Theorem 23, there exists a sequence of orthogonal matrices $Q \in \mathbb{R}^{d\times d}$ such that*

$$\|Q\hat{w}_{\mathrm{LS}} - w_{\mathrm{LS}}\| = O(n^{-1/2}\log n).$$

Proof This is a direct result of Theorem 23 and the preceding Lemmas, once we establish bounds on $\kappa_2(XQ)$ and

$$\nu_{\mathrm{LS}} = \frac{\|XQw_{\mathrm{LS}}\|}{\lambda_d(XQ)\|w_{\mathrm{LS}}\|}.$$

By Lemma 14, we have $C_1\sqrt{n} \geq \lambda_1(XQ) \geq \lambda_d(XQ) \geq C_2\sqrt{n}$, and it follows immediately that $\kappa_2(XQ) \leq C$ eventually. Since $\|XQw_{\mathrm{LS}}\|/\|w_{\mathrm{LS}}\| \leq \|XQ\| \leq \sqrt{n}$, we also have $\nu_{\mathrm{LS}} \leq C$ eventually.

By Lemma 24, we are assured that Theorem 23 applies eventually. Lemmas 24 and 25 ensure that the each of $(\cos\theta_{\mathrm{LS}})^{-1}$ and $\tan\theta_{\mathrm{LS}}$ are bounded by constants eventually. Thus, using Lemma 24 to bound $\|\hat{X} - XQ\|/\|XQ\|$, it follows that the right-hand side of Equation 31 is $O(n^{-1/2}\log n)$ and the result follows. ■

We now turn to showing that w_{LS} is close to the true latent position $\bar{w}$. A combination of this result with Lemma 26 will then yield Theorem 7.

Lemma 27 *Let notation be as above and let $\bar{w} \in \mathbb{R}^d$ (fixed) latent position of the out-of-sample vertex, satisfying $0 \leq \bar{w}^T x \leq 1$ for all $x \in \operatorname{supp} F$. Then for all but finitely many n,*

$$\|w_{\mathrm{LS}} - \bar{w}\| \leq \frac{C\log n}{\sqrt{n}}.$$

Proof Define $\vec{r} = \vec{a} - X\bar{w}$. As noted previously, by definition of w_{LS}, we have

$$\|Xw_{\mathrm{LS}} - \vec{a}\|^2 \leq \|X\bar{w} - \vec{a}\|^2 = \|\vec{r}\|^2,$$

whence plugging in $\vec{a} = X\bar{w} + \vec{r}$ yields $\|Xw_{\text{LS}} - X\bar{w} - \vec{r}\|^2 \leq \|\vec{r}\|^2$. Thus,

$$\|Xw_{\text{LS}} - X\bar{w}\|^2 \leq 2\vec{r}^T X(w_{\text{LS}} - \bar{w}). \tag{35}$$

By Lemma 21, X has full column rank eventually, and thus $\|X(w_{\text{LS}} - \bar{w})\| \geq \sigma_d(X)\|w_{\text{LS}} - \bar{w}\|$ eventually, as well. Combining this fact with Equation (35) and making use of the fact that $\sigma_d^2(X) = \sigma_d(P)$, we have

$$\|w_{\text{LS}} - \bar{w}\|^2 \leq \frac{\|X(w_{\text{LS}} - \bar{w})\|^2}{\sigma_d^2(X)} \leq \frac{2\vec{r}^T X(w_{\text{LS}} - \bar{w})}{\sigma_d(P)}.$$

Applying the Cauchy-Schwartz inequality and dividing by $\|w_{\text{LS}} - \bar{w}\|$ (noting that our result holds trivially when $w_{\text{LS}} = \bar{w}$, so we may safely assume that $\|w_{\text{LS}} - \bar{w}\|$ is nonzero)

$$\|w_{\text{LS}} - \bar{w}\| \leq \frac{2\|X^T\vec{r}\|}{\sigma_d(P)}.$$

Thus, it remains for us to show that $\|X^T\vec{r}\|$ grows at a rate at most $O(\sqrt{n}\log^2 n)$, from which Lemma 14 will yield our desired growth rate. Expanding, we have

$$\|X^T\vec{r}\|_2^2 = \sum_{k=1}^{d}\left(\sum_{i=1}^{n}(a_i - X_i^T\bar{w})X_{i,k}\right)^2. \tag{36}$$

Fixing some $k \in [d]$, Hoeffding's inequality implies that with probability at least $1 - O(n^{-2})$, $|\sum_{i=1}^{n}(a_i - X_i^T\bar{w})X_{i,k}| \leq 2\sqrt{n}\log n$. Since d is assumed to be constant in n, a union bound over all $k \in [d]$ implies $\|X^T\vec{r}\|_2^2 \leq 4dn\log^2 n$ with probability at least $1 - O(n^{-2})$. Applying the Borel-Cantelli Theorem and taking square roots completes the proof. ■

Appendix C. Proof of ASE ML-OOS Concentration Inequality

To prove Theorem 8, we will apply a standard argument from convex optimization and use the properties of the set $\widehat{\mathcal{T}}_\epsilon$ to show that

$$\|Q\hat{w}_{\text{ML}} - \bar{w}\| \leq \frac{\|\nabla\hat{\ell}(Q^T\bar{w})\|}{Cn},$$

where $Q \in \mathbb{R}^{d\times d}$ is the orthogonal matrix guaranteed by Lemma 6. This is proven in Lemma 28. We then show in Lemma 29 that

$$\|\nabla\hat{\ell}(Q^T\bar{w})\| = O(\sqrt{n}\log n),$$

which establishes Theorem 8 by the triangle inequality.

Recall the log-likelihood functions

$$\begin{aligned}\ell(w) &= \sum_{i=1}^{n} a_i \log X_i^T w + (1 - a_i)\log(1 - X_i^T w)\\ \hat{\ell}(w) &= \sum_{i=1}^{n} a_i \log \hat{X}_i^T w + (1 - a_i)\log(1 - \hat{X}_i^T w)\end{aligned} \tag{37}$$

and observe that both are convex in their arguments.

Lemma 28 *With notation as above, under the assumptions of Theorem 8, it holds almost surely that for all suitably large n, there exists an orthogonal matrix $Q \in \mathbb{R}^{d\times d}$ satisfying*

$$\|Q\hat{w}_{\mathrm{ML}} - \bar{w}\| \leq \frac{\|\nabla\hat{\ell}(Q^T\bar{w})\|}{Cn}.$$

Proof By a standard argument, we have

$$\begin{aligned}
&\left(\nabla\hat{\ell}(Q^T\bar{w})\right)^T (Q^T\bar{w} - \hat{w}_{\mathrm{ML}}) \\
&\quad = \left(\nabla\hat{\ell}(\hat{w}_{\mathrm{ML}})\right)^T (Q^T\bar{w} - \hat{w}_{\mathrm{ML}}) \\
&\qquad + \int_0^1 (Q^T\bar{w} - \hat{w}_{\mathrm{ML}})^T \nabla^2\hat{\ell}\left(Q^T\bar{w} + t(Q^T\bar{w} - \hat{w}_{\mathrm{ML}})\right)(Q^T\bar{w} - \hat{w}_{\mathrm{ML}})dt \\
&\quad \geq \|\bar{w} - Q\hat{w}_{\mathrm{ML}}\|^2 \min_{w\in\widehat{\mathcal{T}}_\epsilon} \lambda_{\min}\left(\nabla^2\hat{\ell}(w)\right).
\end{aligned}$$

Rearranging and applying the Cauchy-Schwarz inequality implies

$$\|\bar{w} - Q\hat{w}_{\mathrm{ML}}\| \leq \frac{\|\nabla\hat{\ell}(Q^T\bar{w})\|}{|\lambda_{\min}\left(\nabla^2\hat{\ell}(w)\right)|}.$$

The constraint that $w \in \widehat{\mathcal{T}}_\epsilon$ implies that for suitably large n,

$$\min_{w\in\widehat{\mathcal{T}}_\epsilon} \lambda_{\min}\left(\nabla^2\hat{\ell}(w)\right) \geq Cn,$$

with $C > 0$ depending on ϵ and F but not on n, where we have used our assumption on the existence of $\eta > 0$ to apply Lemma 6, which ensures that $\{\hat{X}_i\}_{i=1}^n$ are uniformly close to $\operatorname{supp} F$. We conclude that eventually,

$$\|\bar{w} - Q\hat{w}_{\mathrm{ML}}\| \leq \frac{\|\nabla\hat{\ell}(Q^T\bar{w})\|}{Cn},$$

completing the proof. ■

Lemma 29 *With notation as above, under the assumptions of Theorem 8,*

$$\|\nabla\hat{\ell}(Q^T\bar{w})\| = O(\sqrt{n}\log n).$$

Proof By the triangle inequality,

$$\|\nabla\hat{\ell}(Q^T\bar{w})\| \leq \|\nabla\ell(\bar{w})\| + \|\nabla\hat{\ell}(Q^T\bar{w}) - \nabla\ell(\bar{w})\|. \tag{38}$$

We will show that both terms on the right hand side of (38) are $O(\sqrt{n}\log^{1/2} n)$.

Fix $k \in [d]$. By our boundedness assumption on $\bar{w}$ and $\operatorname{supp} F$, as well as the fact that $\bar{w}, X_1, X_2, \ldots, X_n \in \operatorname{supp} F$,

$$(\nabla\ell(\bar{w}))_k = \sum_{i=1}^n \left(\frac{a_i}{X_i^T\bar{w}} - \frac{1-a_i}{1-X_i^T\bar{w}}\right) X_{i,k} = \sum_{i=1}^n \frac{(a_i - X_i^T\bar{w})X_{i,k}}{X_i^T\bar{w}(1-X_i^T\bar{w})}$$

is a sum of bounded mean-zero random variables. Applying Hoeffding's inequality,

$$\Pr\left[\,|(\nabla\ell(\bar{w}))_k| \geq t\right] \leq 2\exp\left\{\frac{-2t^2}{Cn}\right\}$$

for some constant $C > 0$ depending on F and $\bar{w}$ but not n. Choosing $t = \sqrt{Cn}\log^{1/2} n$, we have $(\nabla\ell(\bar{w}))_k \geq \sqrt{Cn}\log^{1/2} n$ with probability at most $O(n^{-2})$. A union bound over all $k \in [d]$, implies that with probability at least $1 - Cdn^{-2}$,

$$\sum_{k=1}^{d} (\nabla\ell(\bar{w}))_k^2 \leq dCn\log n,$$

and the Borel-Cantelli Lemma implies $\|\nabla\ell(\bar{w})\| = O(\sqrt{n}\log^{1/2} n)$ after taking square roots.

Turning to the second term on the right hand side of (38), fixing $k \in [d]$, we have

$$\left(\nabla\hat{\ell}(Q^T\bar{w}) - \nabla\ell(\bar{w})\right)_k = \sum_{i=1}^{n} \frac{(a_i - \hat{X}_i^T Q^T\bar{w})\hat{X}_{i,k}}{\hat{X}_i^T Q^T\bar{w}(1 - \hat{X}_i^T Q^T\bar{w})} - \sum_{i=1}^{n} \frac{(a_i - X_i^T\bar{w})X_{i,k}}{X_i^T\bar{w}(1 - X_i^T\bar{w})}.$$

Taking expectation conditional on A and X, the second sum has expectation 0, and

$$\mathbb{E}\left[\left(\nabla\hat{\ell}(Q^T\bar{w}) - \nabla\ell(\bar{w})\right)_k \middle| A, X\right] = \sum_{i=1}^{n} \frac{\left((Q\hat{X}_i) - X_i\right)^T\bar{w}}{(Q\hat{X}_i)^T\bar{w}(1 - (Q\hat{X}_i)^T\bar{w})}\hat{X}_{i,k}.$$

By Lemma 6 and our boundedness assumptions on $\bar{w}$ and $\operatorname{supp} F$, the denominators in this sum are uniformly bounded away from zero over almost all sequences of (A, X). Lemma 6 also bounds the numerators in this sum uniformly by $O(n^{-1/2}\log n)$, and it follows that

$$\mathbb{E}\left[\left(\nabla\hat{\ell}(Q^T\bar{w}) - \nabla\ell(\bar{w})\right)_k \middle| A, X\right] = O(\sqrt{n}\log n). \tag{39}$$

Our proof will be complete if we can show that

$$\left(\nabla\hat{\ell}(Q^T\bar{w}) - \nabla\ell(\bar{w})\right)_k - \mathbb{E}\left[\left(\nabla\hat{\ell}(Q^T\bar{w}) - \nabla\ell(\bar{w})\right)_k \middle| A, X\right]$$

concentrates at the same rate. Toward this end, for ease of notation, for each $i \in [n]$ define $p_i = X_i^T\bar{w}$ and $\hat{p}_i = \hat{X}_i^T\bar{w}$. Then

$$\begin{aligned}
&\left(\nabla\hat{\ell}(Q^T\bar{w}) - \nabla\ell(\bar{w})\right)_k - \mathbb{E}\left[\left(\nabla\hat{\ell}(Q^T\bar{w}) - \nabla\ell(\bar{w})\right)_k \middle| A, X\right] \\
&\qquad = \sum_{i=1}^{n}\left[\frac{(a_i - \hat{p}_i)\hat{X}_{i,k}}{\hat{p}_i(1-\hat{p}_i)} - \frac{(a_i - p_i)X_{i,k}}{p_i(1-p_i)} - \frac{(p_i - \hat{p}_i)\hat{X}_{i,k}}{\hat{p}_i(1-\hat{p}_i)}\right] \\
&\qquad = \sum_{i=1}^{n}(a_i - p_i)\left(\frac{\hat{X}_{i,k}}{\hat{p}_i(1-\hat{p}_i)} - \frac{X_{i,k}}{p_i(1-p_i)}\right).
\end{aligned}$$

Conditional on (A, X), this is a sum of n independent zero-mean random vectors, with the i-th summand bounded by

$$\left|(a_i - p_i)\left(\frac{\hat{X}_{i,k}}{\hat{p}_i(1-\hat{p}_i)} - \frac{X_{i,k}}{p_i(1-p_i)}\right)\right| \leq \left|\frac{\hat{X}_{i,k}}{\hat{p}_i(1-\hat{p}_i)} - \frac{X_{i,k}}{p_i(1-p_i)}\right|$$

since $|a_i - p_i| \leq 1$. Let M_i denote this bound for each $i \in [n]$. Let $s > 0$ be a value which we will specify below, and let B_n denote the event that

$$\left| \left(\nabla \hat{\ell}(Q^T \bar{w}) - \nabla \ell(\bar{w}) \right)_k - \mathbb{E}\left[\left(\nabla \hat{\ell}(Q^T \bar{w}) - \nabla \ell(\bar{w}) \right)_k \middle| A, X \right] \right| > s.$$

Hoeffding's inequality conditional on A, X implies that

$$\Pr\left[B_n \mid A, X\right] \leq 2 \exp\left\{ \frac{-s^2}{2\sum_{i=1}^n M_i^2} \right\}.$$

By definition of M_i, we have

$$\begin{aligned} M_i &= \left| \frac{\hat{X}_{i,k}}{\hat{p}_i(1-\hat{p}_i)} - \frac{X_{i,k}}{p_i(1-p_i)} \right| \\ &\leq \frac{|\hat{X}_{i,k} - X_{i,k}|}{p_i(1-p_i)} + \left| \frac{1}{p_i(1-p_i)} - \frac{1}{\hat{p}_i(1-\hat{p}_i)} \right| |X_{i,k}| \\ &\leq \frac{O(n^{-1/2}\log n)}{p_i(1-p_i)} + \frac{|p_i - \hat{p}_i|(1-p_i) + p_i|p_i - \hat{p}_i|}{p_i(1-p_i)\hat{p}_i(1-\hat{p}_i)}, \end{aligned}$$

where the first inequality follows from the triangle inequality, and the second inequality follows from Lemma 6 and the fact that $\|X_i\| \leq 1$ by definition of F being an inner product distribution. Lemma 6 implies that $|\hat{p}_i - p_i| = O(n^{-1/2}\log n)$, since $\|\bar{w}\|$ is bounded by assumption. Our boundedness assumptions on $\bar{w}$ and the support of F, along with yet another application of Lemma 6, imply that both denominators are bounded away from 0 eventually. Thus, uniformly over all $i \in [n]$, $M_i = O(n^{-1/2}\log n)$, so that $\sum_{i=1}^n M_i^2 = O(\log^2 n)$, and integrating with respect to (A, X) implies that

$$\Pr\left[B_n \mid A, X\right] \leq 2 \exp\left\{ \frac{-Cs^2}{\log^2 n} \right\}.$$

Taking $s = C\log^{3/2} n$ for suitably large constant C and applying the Borel-Cantelli Lemma ensures that B_n occurs eventually, and we have that

$$\left(\nabla \hat{\ell}(Q^T \bar{w}) - \nabla \ell(\bar{w}) \right)_k - \mathbb{E}\left[\left(\nabla \hat{\ell}(Q^T \bar{w}) - \nabla \ell(\bar{w}) \right)_k \middle| A, X \right] = O(\log^{3/2} n).$$

Combining this with Equation (39), we conclude that

$$\left(\nabla \hat{\ell}(Q^T \bar{w}) - \nabla \ell(\bar{w}) \right)_k = O(\sqrt{n}\log n).$$

Since d is assumed constant, this rate holds uniformly over all $k \in [d]$, and we conclude that

$$\|\nabla \hat{\ell}(Q^T \bar{w}) - \nabla \ell(\bar{w})\| = O(\sqrt{n}\log n),$$

completing the proof. ■

Appendix D. Proof of LSE LS-OOS Concentration Inequality

Here we provide a proof of Theorem 9. The argument proceeds similarly to the proof of Theorem 7 in Appendix B above. Recall that $\check{w}_{\mathrm{LS}} \in \mathbb{R}^d$ denotes the least-squares OOS extension, given by the solution to

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left(\frac{a_i}{d_v^{1/2} d_i^{1/2}} - \check{X}_i^T w \right)^2,$$

where $\check{X}_i \in \mathbb{R}^d$ is the LSE estimate of the Laplacian spectral embedding of the true latent position of the i-th vertex and d_i denotes the degree of vertex i for $i \in [n] \cup \{v\}$. We define $\tilde{w}_{\mathrm{LS}} \in \mathbb{R}^d$ to be the least-squares OOS extension if we had access to the true latent positions. That is, $\tilde{w}_{\mathrm{LS}}$ is the solution to the least-squares problem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left(\frac{a_i}{d_v^{1/2} d_i^{1/2}} - \tilde{X}_i^T w \right)^2.$$

Letting $\tilde{Q} \in \mathbb{R}^{d \times d}$ denote the orthogonal matrix guaranteed by Lemma 6, our proof of Theorem 9 will proceed by showing that both $\|\tilde{w}_{\mathrm{LS}} - \tilde{w}\|$ and $\|\check{w}_{\mathrm{LS}} - \tilde{Q}^T \tilde{w}_{\mathrm{LS}}\|$ are $O(n^{-1} \log^{1/2} n)$, after which the triangle inequality will yield our desired result.

Lemma 30 *With notation as above,*

$$\|\tilde{w}_{\mathrm{LS}} - \tilde{w}\| = O(n^{-1} \log^{1/2} n).$$

Proof Recall that $D \in \mathbb{R}^{n \times n}$ is the diagonal matrix of in-sample vertex degrees and $d_v = \sum_{i=1}^n a_i$ denotes the degree of the out-of-sample vertex v. Define $\vec{b} = d_v^{-1/2} D^{-1/2} \vec{a}$, and let $\vec{z} = \vec{b} - \tilde{X}\tilde{w}$. By definition of $\tilde{w}_{\mathrm{LS}}$ as a least squares solution, we have

$$\|\tilde{X}\tilde{w}_{\mathrm{LS}} - \vec{b}\| \leq \|\vec{z}\|.$$

Substituting $\vec{b} = \vec{z} + \tilde{X}\tilde{w}$, expanding the squares of both sizes, rearranging, and applying the Cauchy-Schwarz inequality,

$$\|\tilde{X}(\tilde{w}_{\mathrm{LS}} - \tilde{w})\|^2 \leq 2\vec{z}^T \tilde{X}(\tilde{w}_{\mathrm{LS}} - \tilde{w}) \leq 2\|X^T \vec{z}\| \|\tilde{w}_{\mathrm{LS}} - \tilde{w}\| \tag{40}$$

By Lemma 21, $\tilde{X}$ is full rank eventually, and therefore

$$\|\tilde{X}(\tilde{w}_{\mathrm{LS}} - \tilde{w})\| \geq \sigma_d(\tilde{X}) \|\tilde{w}_{\mathrm{LS}} - \tilde{w}\| \text{ eventually.}$$

Combining this with (40),

$$\|\tilde{w}_{\mathrm{LS}} - \tilde{w}\|^2 \leq \frac{2\|\tilde{X}^T \vec{z}\| \|\tilde{w}_{\mathrm{LS}} - \tilde{w}\|}{\sigma_d^2(\tilde{X})} \text{ eventually.}$$

In the event that $\tilde{w}_{\mathrm{LS}} = \tilde{w}$, our desired bound holds trivially, so we may safely divide through by $\|\tilde{w}_{\mathrm{LS}} - \tilde{w}\|$ to write

$$\|\tilde{w}_{\mathrm{LS}} - \tilde{w}\| \leq \frac{2\|\tilde{X}^T \vec{z}\|}{\sigma_d^2(\tilde{X})} \text{ eventually.}$$

Lemma 19 implies that $\sigma_d^2(\tilde{X}) = \Theta(1)$, so our proof will be complete if we can bound the growth of $\|\tilde{X}^T \vec{z}\|$. We have

$$\|\tilde{X}^T \vec{z}\|^2 = \sum_{k=1}^{d} \left(\sum_{i=1}^{n} z_i \tilde{X}_{i,k} \right)^2 = \sum_{k=1}^{d} Y_k^2,$$

where $Y_k = \sum_{i=1}^{n} z_i \tilde{X}_{i,k}$. Fixing some $k \in [d]$,

$$Y_k = \sum_{i=1}^{n} \left(\frac{X_i^T \bar{w}}{\sqrt{t_i}\sqrt{n\mu^T \bar{w}}} - \frac{a_i}{\sqrt{d_i}\sqrt{d_v}} \right) \frac{X_{i,k}}{\sqrt{t_i}}.$$

Adding and subtracting appropriate quantities,

$$Y_k = \sum_{i=1}^{n} \frac{(X_i^T \bar{w} - a_i)}{t_i \sqrt{n\mu^T \bar{w}}} X_{i,k} + \sum_{i=1}^{n} \frac{a_i X_{i,k}}{\sqrt{t_i}} \left(\frac{1}{\sqrt{t_i}\sqrt{n\mu^T \bar{w}}} - \frac{1}{\sqrt{d_i}\sqrt{d_v}} \right). \tag{41}$$

Conditional on X, the first term is a sum of independent mean-zero random variables, with

$$\frac{(X_i^T \bar{w} - a_i) X_{i,k}}{t_i \sqrt{n\mu^T \bar{w}}} \in \left[\frac{-1}{t_i \sqrt{n\mu^T \bar{w}}}, \frac{1}{t_i \sqrt{n\mu^T \bar{w}}} \right] \text{ almost surely}$$

for each $i \in [n]$. Let G_n denote the event that

$$\left| \sum_{i=1}^{n} \frac{(X_i^T \bar{w} - a_i)}{t_i \sqrt{n\mu^T \bar{w}}} X_{i,k} \right| > s,$$

where $s = s_n > 0$ will be specified below. Conditional Hoeffding's inequality yields

$$\Pr\left[B_n \mid X\right] \leq 2 \exp\left\{ \frac{-n\mu^T \bar{w} s^2}{\sum_{i=1}^{n} t_i^{-2}} \right\}$$

Let B_n denote the event that $\min_i t_i \geq Cn$ for some suitably-chosen constant $C > 0$. Lemma 18 ensures that $\Pr[B_n^c] = O(n^{-2})$, and integrating with respect to $X \in \mathbb{R}^{n \times d}$ yields

$$\Pr[G_n] \leq \Pr[G_n \mid B_n] + \Pr[B_n^c] \leq 2 \exp\left\{ -Cn^2 \mu^T \bar{w} s^2 \right\} + O(n^{-2}).$$

Taking $s = Cn^{-1} \log^{1/2} n$ for $C > 0$ suitably large ensures that both terms on the right-hand side are $O(n^{-2})$, and we have

$$\left| \sum_{i=1}^{n} \frac{(X_i^T \bar{w} - a_i) X_{i,k}}{\sqrt{t_i}\sqrt{t_v}\sqrt{n\mu^T \bar{w}}} \right| = O(n^{-1} \log^{1/2} n). \tag{42}$$

Lemma 18 similarly bounds the second sum in (41):

$$\sum_{i=1}^{n} \frac{a_i X_{i,k}}{\sqrt{t_i}} \left(\frac{1}{\sqrt{t_i}\sqrt{n\mu^T \bar{w}}} - \frac{1}{\sqrt{d_i}\sqrt{d_v}} \right) \leq \frac{C}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{1}{\sqrt{t_i}\sqrt{n\mu^T \bar{w}}} - \frac{1}{\sqrt{d_i}\sqrt{d_v}} \right) a_i X_{i,k}. \tag{43}$$

Adding and subtracting appropriate quantities, the sum becomes

$$\sum_{i=1}^{n}\left(\frac{1}{\sqrt{t_i}\sqrt{n\mu^T\bar{w}}}-\frac{1}{\sqrt{d_i}\sqrt{d_v}}\right)a_iX_{i,k}$$
$$=\sum_{i=1}^{n}\frac{a_iX_{i,k}}{\sqrt{t_i}}\left(\frac{1}{\sqrt{n\mu^T\bar{w}}}-\frac{1}{\sqrt{d_v}}\right)+\sum_{i=1}^{n}\frac{a_iX_{i,k}}{\sqrt{d_v}}\left(\frac{1}{\sqrt{t_i}}-\frac{1}{\sqrt{d_i}}\right),$$

and several applications of Lemma 18 yields that

$$\sum_{i=1}^{n}\left(\frac{1}{\sqrt{t_i}\sqrt{n\mu^T\bar{w}}}-\frac{1}{\sqrt{d_i}\sqrt{d_v}}\right)a_iX_{i,k}=O(n^{-1/2}\log^{1/2}n),$$

whence, applying this to Equation (43), we have

$$\sum_{i=1}^{n}\frac{a_iX_{i,k}}{\sqrt{t_i}}\left(\frac{1}{\sqrt{t_i}\sqrt{n\mu^T\bar{w}}}-\frac{1}{\sqrt{d_i}\sqrt{d_v}}\right)=O(n^{-1}\log^{1/2}n).$$

Applying this and (42) to the right-hand side of (41), $|Y_k|=O(n^{-1}\log^{1/2}n)$ and a union bound over $k\in[d]$ completes the proof. ■

Lemma 31 *With notation as above, there exists a sequence of orthogonal matrices $\tilde{Q}\in\mathbb{R}^{d\times d}$ such that*

$$\|\tilde{Q}\check{w}_{\mathrm{LS}}-\tilde{w}_{\mathrm{LS}}\|=O(n^{-1}\log^{1/2}n).$$

Proof Recall from above our definition $\vec{b}=d_v^{-1/2}D^{-1/2}\vec{a}$, where d_v is the degree of the out-of-sample vertex and $D\in\mathbb{R}^{n\times n}$ is the diagonal matrix of in-sample vertex degrees, and note that $\check{w}_{\mathrm{LS}}=(\check{X}^T\check{X})^{-1}\check{X}^T\vec{b}$. Our main tool, as in Section B, is Theorem 5.3.1 from Golub and Van Loan (2012), quoted above as Theorem 23. Applying that theorem, we have that so long as $\vec{b},\vec{b}-\tilde{X}\tilde{w}_{\mathrm{LS}}$ and $\tilde{w}_{\mathrm{LS}}$ are all non-zero,

$$\frac{\|\check{w}_{\mathrm{LS}}-\tilde{Q}^T\tilde{w}_{\mathrm{LS}}\|}{\|\tilde{Q}^T\tilde{w}_{\mathrm{LS}}\|}\leq\frac{\|\check{X}-\tilde{X}\tilde{Q}\|}{\|\tilde{X}\tilde{Q}\|}\left(\frac{\nu_{\mathrm{LS}}}{\cos\theta_{\mathrm{LS}}}+(1+\nu_{\mathrm{LS}}\tan\theta_{\mathrm{LS}})\kappa_2(\tilde{X}\tilde{Q})\right)+C\frac{\|\check{X}-\tilde{X}\tilde{Q}\|^2}{\|\tilde{X}\tilde{Q}\|^2},$$

where $\theta_{\mathrm{LS}}\in(0,\pi/2)$ with

$$\sin\theta_{\mathrm{LS}}=\frac{\|\tilde{r}_{\mathrm{LS}}\|}{\|\vec{b}\|},\quad\text{and}\quad\nu_{\mathrm{LS}}=\frac{\|\tilde{X}\tilde{w}_{\mathrm{LS}}\|}{\sigma_d(\tilde{X}\tilde{Q})\|\tilde{Q}^T\tilde{w}_{\mathrm{LS}}\|}.$$

In order to apply Theorem 23, we must first show that eventually

1. $\|\check{X}-\tilde{X}\tilde{Q}\|<\sigma_d(\tilde{X})$ and
2. the quantities $\vec{b},\vec{b}-\tilde{X}\tilde{w}_{\mathrm{LS}}$, and $\tilde{w}_{\mathrm{LS}}$ are all non-zero.

The first condition holds eventually by Lemma 19 and the fact that, using the relations between the spectral, Frobenius and $(2,\infty)$-norms,

$$\|\check{X} - \tilde{X}\tilde{Q}\|^2 \leq \|\check{X} - \tilde{X}\tilde{Q}\|_F^2 \leq n\|\check{X}_i - \tilde{Q}\tilde{X}_i\|_{2,\infty}^2 \leq \frac{C\log n}{n}, \tag{44}$$

where the last inequality holds eventually by Lemma 6. Note that application of the Laplacian case of Lemma 6 requires our boundedness assumption on $\operatorname{supp} F$. As in the proof of Lemma 24, it is immediate from the model that condition 2 holds eventually.

Equation (44), along with another application of Lemma 19 to control $\lambda_d(\mathcal{L}(P))$ implies that

$$\frac{\|\check{X} - \tilde{X}\tilde{Q}\|}{\|\tilde{X}\tilde{Q}\|} \leq \frac{C\log^{1/2} n}{\sqrt{n\sigma_d(\mathcal{L}(P))}} \leq \frac{C\log^{1/2} n}{\sqrt{n}} \text{ eventually} \tag{45}$$

Thus, applying Theorem 23, we have

$$\|\check{w}_{\mathrm{LS}} - \tilde{Q}^T\tilde{w}_{\mathrm{LS}}\| \leq \frac{C\|\tilde{w}_{\mathrm{LS}}\|\log^{1/2} n}{\sqrt{n}}\left(\frac{\nu_{\mathrm{LS}}}{\cos\theta_{\mathrm{LS}}} + (1+\nu_{\mathrm{LS}}\tan\theta_{\mathrm{LS}})\kappa_2(\tilde{X}\tilde{Q})\right) + \frac{C\log^2 n}{n^2}. \tag{46}$$

Lemma 19 bounds the condition number $\kappa_2(\tilde{X}\tilde{Q}) = \kappa_2(\tilde{X}) \leq C$, whence

$$\nu_{\mathrm{LS}} = \frac{\|\tilde{X}\tilde{w}_{\mathrm{LS}}\|}{\sigma_d(\tilde{X})\|\tilde{Q}^T\tilde{w}_{\mathrm{LS}}\|} = \frac{\|\tilde{X}\tilde{w}_{\mathrm{LS}}\|}{\sigma_d(\tilde{X})\|\tilde{w}_{\mathrm{LS}}\|} \leq \frac{\|\tilde{X}\|}{\sigma_d(\tilde{X})} = \kappa_2(\tilde{X}) \leq C \text{ eventually.}$$

By the triangle inequality, the definition of $\tilde{w}$ and using Lemma 30 to bound $\|\tilde{w}_{\mathrm{LS}} - \tilde{w}\|$,

$$\|\tilde{w}_{\mathrm{LS}}\| = \left\|\frac{\bar{w}}{\sqrt{n\mu^T\bar{w}}}\right\| + O(n^{-1}\log^{1/2} n) = O(n^{-1/2}) + O(n^{-1}\log^{1/2} n),$$

whence Equation (46) becomes

$$\|\tilde{Q}\check{w}_{\mathrm{LS}} - \tilde{w}_{\mathrm{LS}}\| \leq \frac{C\log^{1/2} n}{n}\left(1 + \frac{1+\sin\theta_{\mathrm{LS}}}{\cos\theta_{\mathrm{LS}}}\right) + \frac{C\log^2 n}{n^2} \text{ eventually.}$$

Thus, to complete the proof, it will suffice to bound $\cos\theta_{\mathrm{LS}}$ away from 0. To do this, we will show by an argument similar to that in Lemma 25 that there exists a constant $\rho \in (0,1)$ such that $\sin\theta_{\mathrm{LS}} \leq 1-\rho$ eventually.

Toward this end, define $\tilde{b} = t_v^{-1/2}T^{-1/2}\vec{a}$, where we remind the reader that $t_v = \sum_{i=1}^n X_i^T\bar{w}$ is the expected degree of the out-of-sample vertex conditioned on the latent positions, and $T \in \mathbb{R}^{n\times n}$ is the diagonal matrix of in-sample vertex expected degrees, i.e., $T_{i,i} = \sum_{j=1}^n X_j^T X_i$. Letting $\tilde{X}^\dagger = (X^TT^{-1}X)^{-1}X^TT^{-1/2}$ denote the pseudoinverse of $\tilde{X}$, (with the inverse existing eventually by Lemma 21), we have

$$\begin{aligned}\sin\theta_{\mathrm{LS}} &= \frac{\|\vec{b} - \tilde{X}\tilde{w}_{\mathrm{LS}}\|}{\|\vec{b}\|} = \frac{\|(I-\tilde{X}\tilde{X}^\dagger)\vec{b}\|}{\|\vec{b}\|} = \frac{\|\tilde{b}\|}{\|\vec{b}\|}\frac{\|(I-\tilde{X}\tilde{X}^\dagger)\vec{b}\|}{\|\tilde{b}\|} \\ &\leq \frac{\|\tilde{b}\|}{\|\vec{b}\|}\left(\frac{\|I-\tilde{X}\tilde{X}^\dagger\|\|\vec{b}-\tilde{b}\|}{\|\tilde{b}\|} + \frac{\|(I-\tilde{X}\tilde{X}^\dagger)\tilde{b}\|}{\|\tilde{b}\|}\right),\end{aligned} \tag{47}$$

where the inequality follows from the triangle inequality and submultiplicativity. By definition of $\vec{b}$ and $\tilde{b}$, we have

$$\frac{\|\vec{b}-\tilde{b}\|}{\|\tilde{b}\|} = \frac{\left\|(d_v^{-1/2}D^{-1/2} - t_v^{-1/2}T^{-1/2})\vec{a}\right\|}{\|t_v^{-1/2}T^{-1/2}\vec{a}\|} \leq \frac{\|d_v^{-1/2}D^{-1/2} - t_v^{-1/2}T^{-1/2}\|}{t_v^{-1/2}/\max_i \sqrt{t_i}},$$

where we have used submultiplicativity to upper bound the numerator, $\|T^{-1/2}\vec{a}\| \geq \|\vec{a}\|/\max_i \sqrt{t_i}$ to lower-bound the denominator, and cancelled the resulting factor of $\|\vec{a}\|$. Cancelling factors of $t_v^{-1/2}$, we have

$$\frac{\|\vec{b}-\tilde{b}\|}{\|\tilde{b}\|} \leq \|t_v^{1/2}d_v^{-1/2}D^{-1/2} - T^{-1/2}\| \max_i \sqrt{t_i}.$$

Lemma 18 implies $\max_i \sqrt{t_i} = O(\sqrt{n})$, and a second application of Lemma 18 implies that $\|t_v^{1/2}d_v^{-1/2}D^{-1/2} - T^{-1/2}\| = O(n^{-1}\log^{1/2} n)$, from which

$$\frac{\|\vec{b}-\tilde{b}\|}{\|\tilde{b}\|} = O(n^{-1/2}\log^{1/2} n), \tag{48}$$

and it follows from the triangle inequality that

$$\frac{\|\tilde{b}\|}{\|\vec{b}\|} \leq \frac{\|\vec{b}\| + \|\tilde{b}-\vec{b}\|}{\|\vec{b}\|} = 1 + O(n^{-1/2}\log^{1/2} n) = O(1). \tag{49}$$

Applying Equations (48) and (49) to Equation (47) and using the bound $\|I - \tilde{X}\tilde{X}^\dagger\| \leq 1$,

$$\sin\theta_{\mathrm{LS}} \leq O\left(\frac{\log^{1/2} n}{\sqrt{n}}\right) + \frac{C\|(I-\tilde{X}\tilde{X}^\dagger)\tilde{b}\|}{\|\tilde{b}\|}. \tag{50}$$

Letting $\mathcal{P}_{\tilde{X}}^{\perp} = (I - \tilde{X}\tilde{X}^\dagger)$ denote the orthogonal projection onto the orthogonal complement of the column space of $\tilde{X} = T^{-1/2}X$, we have, canceling factors of $t_v^{-1/2}$ in the numerator and denominator,

$$\frac{\|(I-\tilde{X}\tilde{X}^\dagger)\tilde{b}\|}{\|\tilde{b}\|} = \frac{\|(I-\tilde{X}\tilde{X}^\dagger)T^{-1/2}\vec{a}\|}{\|T^{-1/2}\vec{a}\|} = \frac{\|\mathcal{P}_{\tilde{X}}^{\perp}T^{-1/2}\vec{a}\|}{\|T^{-1/2}\vec{a}\|} = \frac{\|\mathcal{P}_{\tilde{X}}^{\perp}T^{-1/2}(\vec{a}-X\bar{w})\|}{\|T^{-1/2}\vec{a}\|},$$

where we have used the fact that $\mathcal{P}_{\tilde{X}}^{\perp}T^{-1/2}X\bar{w} = 0$, since $T^{-1/2}X\bar{w} = \tilde{X}\bar{w}$ is in the column space of $\tilde{X}$. Thus, defining $\vec{r} = \vec{a} - X\bar{w}$, we have

$$\frac{\|(I-\tilde{X}\tilde{X}^\dagger)\tilde{b}\|}{\|\tilde{b}\|} = \frac{\|\mathcal{P}_{\tilde{X}}^{\perp}T^{-1/2}\vec{r}\|}{\|T^{-1/2}\vec{a}\|} \leq \frac{\|T^{-1/2}\|\|\vec{r}\|}{\|\vec{a}\|/\max_i \sqrt{t_i}} \leq C\frac{\|\vec{r}\|}{\|\vec{a}\|},$$

where the last inequality follows from the fact that the expected degrees $\{t_i\}_{i=1}^n$ are all of the same order by Lemma 18. The same argument as that given in the proof of Lemma 25

lets us bound $\|\vec{r}\|/\|\vec{a}\|$ by a constant $\rho > 0$ smaller than $1/(2C)$. Applying this to (50), we obtain

$$\sin\theta_{\mathrm{LS}} \leq 1 - \rho + O(n^{-1/2}\log^{1/2} n)$$

It follows that

$$\sin\theta_{\mathrm{LS}} \leq 1 - \frac{\rho}{2} \text{ eventually},$$

i.e., $\sin\theta_{\mathrm{LS}}$ is bounded away from 1, completing the proof. ■

Appendix E. Proof of ASE linear least squares out-of-sample CLT

In this section, we prove Theorem 11, which shows that taking $\{Q_n\}_{n=1}^{\infty}$ to be the sequence of orthogonal d-by-d matrices guaranteed to exist by Lemma 6, the quantity $\sqrt{n}(\hat{w}_{\mathrm{LS}} - Q^T\bar{w})$ is asymptotically multivariate normal. We begin by recalling that

$$\hat{w}_{\mathrm{LS}} = (\hat{X}^T\hat{X})^{-1}\hat{X}^T\vec{a} = \hat{S}^{-1/2}\hat{U}^T\vec{a}.$$

Our proof will consist of writing $\sqrt{n}(\hat{w}_{\mathrm{LS}} - Q^T\bar{w})$ as a sum of two random vectors,

$$\sqrt{n}(\hat{w}_{\mathrm{LS}} - Q^T\bar{w}) = \sqrt{n}\vec{g} + \sqrt{n}\vec{h},$$

and showing that $\sqrt{n}\vec{g}$ converges in law to a normal, while $\sqrt{n}\vec{h}$ converges in probability to 0. The multivariate version of Slutsky's Theorem will then yield the desired result. We begin by showing that $\vec{g} = \sqrt{n}S^{-1/2}U^T(\vec{a} - X\bar{w})$ will suffice. We remind the reader that $\Delta = \mathbb{E}X_1X_1^T \in \mathbb{R}^{d\times d}$ is the second moment matrix of the latent position distribution F.

Lemma 32 *Let F be a d-dimensional inner product distribution, with $(A, X) \sim \mathrm{RDPG}(F, n)$ and let $\bar{w} \in \mathbb{R}^d$ be such that $0 \leq \bar{w}^T x \leq 1$ for all $x \in \operatorname{supp} F$ be the fixed latent position of the out-of-sample vertex. Then*

$$\sqrt{n}S^{-1/2}U^T(\vec{a} - X\bar{w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{F,\bar{w}}),$$

where $\Sigma_{F,\bar{w}} = \Delta^{-1}\mathbb{E}\left[X_1^T\bar{w}(1 - X_1^T\bar{w})X_1X_1^T\right]\Delta^{-1}$.

Proof We begin by observing that since $\bar{w} \in \mathbb{R}^d$ is fixed,

$$n^{-1/2}X^T(\vec{a} - X\bar{w}) = n^{-1/2}\sum_{i=1}^{n}(\vec{a}_i - X_i^T\bar{w})X_i$$

is a scaled sum of of n independent 0-mean d-dimensional random vectors, each with covariance matrix

$$V_{\bar{w}} = \mathbb{E}X_1^T\bar{w}(1 - X_1^T\bar{w})X_1X_1^T \in \mathbb{R}^{d\times d}.$$

The multivariate central limit theorem implies that

$$n^{-1/2}X^T(\vec{a} - X\bar{w})X_i \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_{\bar{w}}).$$

We have $\sqrt{n}S^{-1/2}U^T(\vec{a} - X\bar{w}) = nS^{-1}n^{-1/2}X^T(\vec{a} - X\bar{w})$. By the WLLN, $S/n \xrightarrow{P} \Delta$, and hence by the continuous mapping theorem, $nS^{-1} \xrightarrow{P} \Delta^{-1}$. Thus, the multivariate version of Slutsky's Theorem implies that

$$\sqrt{n}S^{-1/2}U^T(\vec{a} - X\bar{w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Delta^{-1}V_{\bar{w}}\Delta^{-1}),$$

as we set out to show. ■

The following technical lemma will be crucial for proving one of the convergence results required by our main theorem. Its comparative complexity merits stating it here rather than including it in the proof of Theorem 11 below. We remind the reader that $\hat{S}, S \in \mathbb{R}^{d\times d}$ are the diagonal matrices formed by the top d eigenvalues of A and P, repsectively, and $\hat{U}, U \in \mathbb{R}^{n\times d}$ are the matrices whose columns are the corresponding unit eigenvectors.

Lemma 33 *With notation as above,*

$$\sqrt{n}\hat{S}^{-1/2}(\hat{U}^T - \hat{U}^T U U^T)(\vec{a} - X\bar{w}) \xrightarrow{P} 0.$$

Proof For ease of notation, define the vector

$$\vec{z} = (\hat{U}^T - \hat{U}^T U U^T)(\vec{a} - X\bar{w}).$$

Let $\epsilon > 0$ be a constant, and note that for suitably large n,

$$\Pr\left[\sqrt{n}\|\hat{S}^{-1/2}\vec{z}\| > \epsilon\right] \leq \Pr\left[\sqrt{n}\|\hat{S}^{-1/2}\vec{z}\| > C_0 n^{-1/4}\right],$$

where $C_0 > 0$ is a constant that we are free to choose. Define the events

$$E_{1,n} = \{\|\hat{S}^{-1/2}\| \leq C_1 n^{-1/2}\},$$
$$\text{and}$$
$$E_{2,n} = \{\sqrt{n}\|\vec{z}\| \leq C_2 n^{1/4}\},$$

and note that $\Pr\left[\sqrt{n}\|\hat{S}^{-1/2}\vec{z}\| > C_0 n^{-1/4}\right] \leq \Pr\left[(E_{1,n} \cap E_{2,n})^c\right]$ so long as $C_1 C_2 \leq C_0$. Thus, it will suffice for us to show that $\lim_{n\to\infty} \Pr\left[(E_{1,n} \cap E_{2,n})^c\right] \to 0$. The proof of Lemma 14 implies that $\lim_{n\to\infty} \Pr[E_{1,n}^c] = 0$, so our proof will be complete once we show that $\lim_{n\to\infty} \Pr[E_{2,n}^c] = 0$.

Toward this end, define the matrix

$$W = e_n^T \otimes \bar{w} = \begin{bmatrix}\bar{w} & \bar{w} & \dots & \bar{w}\end{bmatrix} \in \mathbb{R}^{d\times n}$$

and let $B \in \mathbb{R}^{n\times n}$ be a random matrix with independent binary entries with $\mathbb{E}B_{i,j} = (XW)_{i,j} = X_i^T\bar{w}$. Define the event

$$E_{3,n} = \{\|(\hat{U}^T - \hat{U}^T U U^T)(B - XW)\|_F^2 \leq C\log^2 n\}.$$

Since $\Pr[E_{2,n}^c] \leq \Pr[E_{2,n}^c \mid E_{3,n}] + \Pr[E_{3,n}^c]$, it will suffice to show that

1. $\lim_{n\to\infty} \Pr\left[E_{3,n}^c\right] = 0$, and
2. $\lim_{n\to\infty} \Pr\left[E_{2,n}^c \mid E_{3,n}\right] = 0$.

By submultiplicativity, we have

$$\|(\hat{U}^T - \hat{U}^T U U^T)(B - XW)\|_F^2 \leq \|\hat{U}^T - \hat{U}^T U U^T\|_F^2 \|B - XW\|^2. \tag{51}$$

Theorem 22 applied to $B - XW$ implies that with probability $1 - O(n^{-2})$,

$$\|B - XW\| \leq C n^{1/2} \log^{1/2} n. \tag{52}$$

Theorem 2 in Yu et al. (2015) guarantees an orthogonal $R^* \in \mathbb{R}^{d\times d}$ such that

$$\|\hat{U} - UR^*\|_F \leq \frac{C\|A - P\|}{\lambda_d(P)} = O\left(\frac{\log^{1/2} n}{\sqrt{n}}\right), \tag{53}$$

where we have used Lemma 14 to lower-bound $\lambda_d(P)$ and bounded $\|A-P\| = O(n^{1/2}\log^{1/2} n)$ by a result in Oliveira (2010). Since $R = \hat{U}^T U$ solves the minimization

$$\min_{R\in\mathbb{R}^{d\times d}} \|\hat{U}^T R - \hat{U}^T U U^T\|_F,$$

Equation (53) implies

$$\|\hat{U}^T - \hat{U}^T U U^T\|_F \leq \|\hat{U}^T - R^* U^T\|_F = O(n^{-1/2}\log^{1/2} n).$$

Plugging this and (52) back into (51), we have that with probability $1 - O(n^{-2})$,

$$\|(\hat{U}^T - \hat{U}^T U U^T)(B - XW)\|_F^2 \leq C \log^2 n \tag{54}$$

which is to say, $\Pr[E_{3,n}^c] = O(n^{-2})$.

It remains to show that $\Pr[E_{2,n}^c \mid E_{3,n}] \to 0$. By construction, the columns of the matrix $(\hat{U}^T - \hat{U}^T U U^T)(B - XW)$ are n independent copies of $\vec{z}$. Using this fact and the conditional Markov inequality, we have

$$\begin{aligned}\Pr[E_{2,n}^c \mid E_{3,n}] &= \Pr[\sqrt{n}\|\vec{z}\| > C_2 n^{1/4} \mid E_{3,n}] \leq \frac{n\mathbb{E}[\|\vec{z}\|^2 \mid E_{3,n}]}{C_2^2 n^{1/2}} \\ &= \frac{\mathbb{E}[\|(\hat{U}^T - \hat{U}^T U U^T)(B - XW)\|_F^2 \mid E_{3,n}]}{C_2^2 n^{1/2}} \leq \frac{C\log^2 n}{n^{1/2}},\end{aligned}$$

where the last inequality follows from the definition of event $E_{3,n}$. This quantity goes to zero in n, thus completing the proof. ■

The following technical lemma will prove useful in our proof of Theorem 11 below. We state it here rather than proving it in-line for the sake of clarity.

Lemma 34 *With notation as above,*

$$\|U^T(\vec{a} - X\bar{w})\| = O(n^{1/2}\log^{1/2} n).$$

Proof For $k \in [d]$ and $i \in [n]$, observe that

$$\left(U^T(\vec{a} - X\bar{w})\right)_{k,i} = \sum_{j=1}^{n} (U)_{j,k}(a_j - X_j^T \bar{w})$$

is a sum of independent 0-mean random variables, and Hoeffding's inequality yields

$$\Pr\left[|U^T(\vec{a} - X\bar{w})|_{k,i} \geq t\right] \leq 2\exp\left\{\frac{-t^2}{2\sum_{j=1}^{n}(U)_{k,j}^2}\right\} = 2\exp\left\{\frac{-t^2}{2}\right\}.$$

Taking $t = C\log^{1/2} n$ for suitably large constant $C > 0$, a union bound over all $k \in [d]$ and $i \in [n]$ followed by the Borel-Cantelli Lemma yields the result. ■

We are now ready to present the proof of Theorem 11.

Proof [Proof of Theorem 11] Let $Q = Q_n \in \mathbb{R}^{d\times d}$ denote the orthogonal matrix guaranteed to exist by Lemma 6. Adding and subtracting appropriate quantities,

$$\begin{aligned}
\sqrt{n}(Q\hat{w}_{\mathrm{LS}} - \bar{w}) &= \sqrt{n}Q\left(\hat{S}^{-1/2}\hat{U}^T\vec{a} - Q^T\bar{w}\right) \\
&= \sqrt{n}S^{-1/2}U^T(\vec{a} - X\bar{w}) \\
&\quad + \sqrt{n}Q\hat{S}^{-1/2}(\hat{U}^T - Q^TU^T)(\vec{a} - X\bar{w}) \\
&\quad + \sqrt{n}Q(\hat{S}^{-1/2}\hat{U}^TX - Q^T)\bar{w} \\
&\quad + \sqrt{n}Q(\hat{S}^{-1/2}Q^T - Q^TS^{-1/2})U^T(\vec{a} - X\bar{w}).
\end{aligned} \tag{55}$$

By Lemma 32, the first of these terms converges in law:

$$\sqrt{n}S^{-1/2}U^T(\vec{a} - X\bar{w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{F,\bar{w}}), \tag{56}$$

where $\Sigma_{F,\bar{w}}$ is as defined in Lemma 32. Thus, by Slutsky's Theorem, our proof will be complete once we show that the remaining terms in Equation (55) go to zero in probability.

Since Q is orthogonal, it suffices to prove that

$$\sqrt{n}\hat{S}^{-1/2}(\hat{U}^T - Q^TU^T)(\vec{a} - X\bar{w}) \xrightarrow{P} 0, \tag{57}$$

$$\sqrt{n}(\hat{S}^{-1/2}\hat{U}^TX - Q^T)\bar{w} \xrightarrow{P} 0, \tag{58}$$

and

$$\sqrt{n}(\hat{S}^{-1/2}Q^T - Q^TS^{-1/2})U^T(\vec{a} - X\bar{w}) \xrightarrow{P} 0. \tag{59}$$

We will address each of these three convergences in order.

To see the convergence in (57), adding and subtracting appropriate quantities gives

$$\begin{aligned}
\sqrt{n}\hat{S}^{-1/2}(\hat{U}^T - Q^TU^T)(\vec{a} - X\bar{w}) &= \sqrt{n}\hat{S}^{-1/2}(\hat{U}^TUU^T - Q^TU^T)(\vec{a} - X\bar{w}) \\
&\quad + \sqrt{n}\hat{S}^{-1/2}(\hat{U}^T - \hat{U}^TUU^T)(\vec{a} - X\bar{w}).
\end{aligned} \tag{60}$$

To bound the first of these two summands, Lemmas 14, 34 and 15 imply

$$\begin{aligned}\|\sqrt{n}\hat{S}^{-1/2}(\hat{U}^T U U^T - Q^T U^T)(\vec{a} - X\bar{w})\| &\leq \sqrt{n}\|\hat{S}^{-1/2}\|\|\hat{U}^T U - Q^T\|\|U^T(\vec{a} - X\bar{w})\|_F \\ &= O(n^{-1/2}\log^{3/2} n).\end{aligned}$$

Lemma 33 shows that the second term in (60) also goes to zero in probability, and Equation (57) follows.

To see (58), note that

$$\begin{aligned}\sqrt{n}(\hat{S}^{-1/2}\hat{U}^T X - Q^T)\bar{w} &= \sqrt{n}\left(\hat{S}^{-1/2}\hat{U}^T U S^{1/2} - Q^T\right)\bar{w} \\ &= \sqrt{n}\hat{S}^{-1/2}\left(\hat{U}^T U - Q^T\right)S^{1/2}\bar{w} + \sqrt{n}\hat{S}^{-1/2}\left(Q^T S^{1/2} - \hat{S}^{1/2}Q^T\right)\bar{w}. \qquad (61)\end{aligned}$$

Submultiplicativity of matrix norms combined with Lemmas 14 and 15 and the fact that $\|\bar{w}\|$ is bounded imply

$$\begin{aligned}\|\sqrt{n}\hat{S}^{-1/2}\left(\hat{U}^T U - Q^T\right)S^{1/2}\bar{w}\| &\leq C\sqrt{n}\|\hat{S}^{-1/2}\|\|\hat{U}^T U - Q^T\|_F\|S^{1/2}\|\|\bar{w}\| \\ &= O(n^{-1/2}\log n).\end{aligned} \qquad (62)$$

Applying Lemma 14 again and taking the Frobenius norm as a trivial upper bound on the spectral norm, Lemma 16 implies

$$\begin{aligned}\|\sqrt{n}\hat{S}^{-1/2}\left(Q^T S^{1/2} - \hat{S}^{1/2}Q^T\right)\bar{w}\| &\leq C\sqrt{n}\|\hat{S}^{-1/2}\|\|Q^T S^{1/2} - \hat{S}^{1/2}Q^T\|\|\bar{w}\| \\ &\leq C\|QS^{1/2} - \hat{S}^{1/2}Q\|,\end{aligned} \qquad (63)$$

where we have used the fact that the spectral norm is preserved by matrix transposition. Adding and subtracting appropriate quantities,

$$QS^{1/2} - \hat{S}^{1/2}Q = (Q - \hat{U}^T U)S^{1/2} + \hat{S}^{1/2}(\hat{U}^T U - Q) + \hat{U}^T U S^{1/2} - \hat{S}^{1/2}\hat{U}^T U.$$

By the triangle inequality and submultiplicativity,

$$\|QS^{1/2} - \hat{S}^{1/2}Q\| \leq \left(\|S^{1/2}\| + \|\hat{S}^{1/2}\|\right)\|\hat{U}^T U - Q\| + \|\hat{U}^T U S^{1/2} - \hat{S}^{1/2}\hat{U}^T U\|. \qquad (64)$$

Lemmas 14 and 15 bound the first term as $O(n^{-1/2}\log n)$, and the second term is bounded by Lemma 17, and thus Equation (63) is bounded as

$$\|\sqrt{n}\hat{S}^{-1/2}\left(Q^T S^{1/2} - \hat{S}^{1/2}Q^T\right)\bar{w}\| = O(n^{-1/2}\log n).$$

Applying this and Equation (62) to Equation (61) proves (58) by the triangle inequality.

Finally, to prove (59), note that

$$\|\sqrt{n}(\hat{S}^{-1/2}Q^T - Q^T S^{-1/2})U^T(\vec{a} - X\bar{w})\| \leq \sqrt{n}\|\hat{S}^{-1/2}Q^T - Q^T S^{-1/2}\|\|U^T(\vec{a} - X\bar{w}\|_F.$$

Lemmas 17 and 34 and an argument similar to the bound in Equation (64) imply that

$$\|\sqrt{n}(\hat{S}^{-1/2}Q^T - Q^T S^{-1/2})U^T(\vec{a} - X\bar{w})\| = O(n^{-1/2}\log^{3/2} n),$$

which completes the proof. ■

Appendix F. Proof of LSE linear least squares out-of-sample CLT

In this section, we prove Theorem 13, which shows that the least-squares out-of-sample extension for the Laplacian spectral embedding is, in the large-n limit, normally distributed about the true embedding $\tilde{w} = \bar{w}/\sqrt{n\mu^T\bar{w}}$, after appropriate rescaling. We remind the reader that $\vec{a} \in \mathbb{R}^n$ denotes the vector of edges between the out-of-sample vertex v and the in-sample vertices $V = [n]$ and $D \in \mathbb{R}^n$ is the diagonal matrix of in-sample node degrees, so that $D_{i,i} = d_i = \sum_{j=1}^n A_{i,j}$. Below, we will also need to define the matrix

$$T = \operatorname{diag}(t_1, t_2, \ldots, t_n) \in \mathbb{R}^{n\times n}, \quad t_i = \sum_{j=1}^n X_j^T X_i,$$

the matrix of in-sample expected degrees conditioned on the latent positions. Analogously, we denote the out-of-sample vertex degree $d_v = \sum_{j=1}^n a_j$, and its expectation $t_v = \sum_{j=1}^n X_j^T \bar{w}$. Recall that the LSE least-squares out-of-sample extension is given by

$$\check{w}_{\mathrm{LS}} = (\check{X}^T\check{X})^{-1}\check{X}^T D^{-1/2}\frac{\vec{a}}{\sqrt{d_v}}.$$

Our aim is to prove that for a suitably-chosen sequence of orthogonal matrices $\tilde{Q} \in \mathbb{R}^{d\times d}$,

$$n(\tilde{Q}\check{w}_{\mathrm{LS}} - \tilde{w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tilde{\Sigma}_{F,\bar{w}}),$$

where $\tilde{\Sigma}_{F,\bar{w}}$ depends only on the latent position distribution F and the true out-of-sample latent position $\bar{w} \in \operatorname{supp} F$, and is given by

$$\tilde{\Sigma}_{F,\bar{w}} = \mathbb{E}\left[\frac{X_j^T\bar{w}(1 - X_j^T\bar{w})}{\mu^T\bar{w}}\left(\frac{\tilde{\Delta}X_j}{X_j^T\mu} - \frac{\bar{w}}{2\mu^T\bar{w}}\right)\left(\frac{\tilde{\Delta}X_j}{X_j^T\mu} - \frac{\bar{w}}{2\mu^T\bar{w}}\right)^T\right] \in \mathbb{R}^{d\times d},$$

where $\tilde{\Delta} = \mathbb{E}X_1X_1^T/(X_1^T\mu)$ with $\mu = \mathbb{E}X_1$ is the mean of F.

Proof [Proof of Theorem 13] Take $\tilde{Q} \in \mathbb{R}^{d\times d}$ to be the matrix guaranteed by Lemma 6. Similarly to the proof of Theorem 11, our proof will proceed by writing $n(\check{w}_{\mathrm{LS}} - \tilde{Q}\tilde{w})$ as

$$n(\tilde{Q}\check{w}_{\mathrm{LS}} - \tilde{w}) = n\vec{g}_n + n\vec{h}_n,$$

where $n\vec{h}_n \xrightarrow{P} 0$ and $n\vec{g}_n$ converges in law to our desired normal distribution, whence Slutsky's Theorem will yield the result. We begin by writing

$$n(\check{w}_{\mathrm{LS}} - \tilde{Q}^T\tilde{w}) = n(\check{X}^T\check{X})^{-1}\frac{\check{X}^T D^{-1/2}\vec{a}}{\sqrt{d_v}} - n\check{U}^T\tilde{U}\tilde{w} - n(\tilde{Q}^T - \check{U}^T\tilde{U})\tilde{w}. \tag{65}$$

By submultiplicativity of the spectral norm, Lemma 16 and the definition of $\tilde{w} = \bar{w}/\sqrt{n\mu^T\bar{w}}$,

$$\|(\tilde{Q}^T - \check{U}^T\tilde{U})\tilde{w}\| \le \|\tilde{Q}^T - \check{U}^T\tilde{U}\|\|\tilde{w}\| \le \frac{C\|\bar{w}\|}{n^{3/2}}.$$

Applying this to Equation (65) and using the fact that $\|\bar{w}\|$ is bounded, we have

$$n(\check{w}_{\mathrm{LS}} - \tilde{Q}^T\tilde{w}) = n(\check{X}^T\check{X})^{-1}\frac{\check{X}^T D^{-1/2}\vec{a}}{\sqrt{d_v}} - n\check{U}^T\tilde{U}\tilde{w} + O(n^{-1/2}). \tag{66}$$

Adding and subtracting quantities,

$$\check{U}^T\tilde{U}\tilde{w} = \check{S}^{-1/2}\check{U}^T\tilde{U}\tilde{S}^{1/2}\tilde{w} - (\check{S}^{-1/2}\check{U}^T\tilde{U}\tilde{S}^{1/2} - \check{U}^T\tilde{U})\tilde{w}. \tag{67}$$

By Lemma 17,

$$\left\|\check{U}^T\tilde{U}\tilde{S}^{1/2} - \check{S}^{1/2}\check{U}^T\tilde{U}\right\| = O(n^{-1}),$$

so that, applying submultiplicativity followed by Lemmas 19 and 17,

$$\left\|(\check{S}^{-1/2}\check{U}^T\tilde{U}\tilde{S}^{1/2} - \check{U}^T\tilde{U})\tilde{w}\right\| \leq \|\check{S}^{-1/2}\|\|\check{U}^T\tilde{U}\tilde{S}^{1/2} - \check{S}^{1/2}\check{U}^T\tilde{U}\|\|\tilde{w}\| = O(n^{-3/2}).$$

Plugging this into Equation (67), we have shown that

$$n\check{U}^T\tilde{U}\tilde{w} = n\check{S}^{-1/2}\check{U}^T\tilde{U}\tilde{S}^{1/2}\tilde{w} + O(n^{-1/2}),$$

and plugging this, in turn, into Equation (66), we have

$$\begin{aligned} n(\check{w}_{\mathrm{LS}} - \tilde{Q}^T\tilde{w}) &= n(\check{X}^T\check{X})^{-1}\frac{\check{X}^T D^{-1/2}\vec{a}}{\sqrt{d_v}} - n\check{S}^{-1/2}\check{U}^T\tilde{U}\tilde{S}^{1/2}\tilde{w} + O(n^{-1/2}) \\ &= n(\check{X}^T\check{X})^{-1}\check{X}^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}} - \tilde{X}\tilde{w}\right) + O(n^{-1/2}), \end{aligned}$$

where the second equality follows from the definitions of $\check{X}$ and $\tilde{X}$ and $\check{X}^T\check{X} = \check{S}$. Again adding and subtracting quantities, we have

$$\begin{aligned} n(\check{w}_{\mathrm{LS}} - \tilde{Q}^T\tilde{w}) = n(\check{X}^T\check{X})^{-1}\tilde{Q}^T\tilde{X}^T&\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}} - \tilde{X}\tilde{w}\right) \\ &+ n(\check{X}^T\check{X})^{-1}(\check{X} - \tilde{X}\tilde{Q})^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}} - \tilde{X}\tilde{w}\right) + O(n^{-1/2}). \end{aligned} \tag{68}$$

Expanding the second term on the right-hand side,

$$\begin{aligned} (\check{X} - \tilde{X}\tilde{Q})^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}} - \tilde{X}\tilde{w}\right) &= \sum_{j=1}^{n}\left(\frac{a_j}{\sqrt{d_j d_v}} - \frac{X_j^T\bar{w}}{\sqrt{t_j n\mu^T\bar{w}}}\right)(\check{X}_j - \tilde{Q}^T\tilde{X}_j) \\ = \sum_{j=1}^{n}\frac{a_j - X_j^T\bar{w}}{\sqrt{d_j d_v}}(\check{X}_j - \tilde{Q}^T\tilde{X}_j) &+ \sum_{j=1}^{n}\left(\frac{1}{\sqrt{t_j n\mu^T\bar{w}}} - \frac{1}{\sqrt{d_j d_v}}\right)X_j^T\bar{w}(\check{X}_j - \tilde{Q}^T\tilde{X}_j). \end{aligned}$$

Recalling that $\vec{a}$ is independent of A conditioned on X and that $\mathbb{E}[a_j \mid X_j] = X_j^T\bar{w}$, the first of these two summations is a sum of independent mean-zero random variables, and an application of Hoeffding's inequality along with Lemmas 6 and 18 yields

$$\sum_{j=1}^{n}\frac{a_j - X_j^T\bar{w}}{\sqrt{d_j d_v}}(\check{X}_j - \tilde{Q}^T\tilde{X}_j) = O(n^{-3/2}\log n).$$

Again applying Lemmas 6 and 18 (and using our boundedness assumption required by the Laplacian case of Lemma 6),

$$\begin{aligned}
&\sum_{j=1}^{n}\left(\frac{1}{\sqrt{t_j n\mu^T\bar{w}}}-\frac{1}{\sqrt{d_j d_v}}\right)X_j^T\bar{w}(\check{X}_j-\tilde{Q}^T\tilde{X}_j)\\
&\quad=\sum_{j=1}^{n}\left(\frac{1}{\sqrt{n\mu^T\bar{w}}}-\frac{1}{\sqrt{d_v}}\right)\frac{X_j^T\bar{w}}{\sqrt{t_j}}\left(\check{X}_j-\tilde{Q}^T\tilde{X}_j\right)+\sum_{j=1}^{n}\left(\frac{1}{\sqrt{t_j}}-\frac{1}{\sqrt{d_j}}\right)\frac{X_j^T\bar{w}}{\sqrt{d_v}}\left(\check{X}_j-\tilde{Q}^T\tilde{X}_j\right)\\
&\quad=O(n^{-3/2}\log n)
\end{aligned}$$

Thus, the above two displays imply that

$$(\check{X}-\tilde{X}\tilde{Q})^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}}-\tilde{X}\tilde{w}\right)=O(n^{-3/2}\log n).$$

Recalling that $\check{S}=\check{X}^T\check{X}$, Lemmas 20 and 21 imply that $\check{S}$ is invertible eventually, and $\|(\check{X}^T\check{X})^{-1}\|=\Theta(1)$. Equation (68) thus becomes

$$n(\check{w}_{\mathrm{LS}}-\tilde{Q}^T\tilde{w})=n\check{S}^{-1}\tilde{Q}^T\tilde{X}^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}}-\tilde{X}\tilde{w}\right)+\tilde{O}(n^{-1/2}),$$

and multiplying through by $\tilde{Q}$ yields

$$n(\tilde{Q}\check{w}_{\mathrm{LS}}-\tilde{w})=n\tilde{Q}\check{S}^{-1}\tilde{Q}^T\tilde{X}^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}}-\tilde{X}\tilde{w}\right)+\tilde{O}(n^{-1/2}).$$

Lemma 20 and the continuity of the inverse imply that

$$\tilde{Q}\check{S}^{-1}\tilde{Q}^T\xrightarrow{P}\tilde{\Delta}^{-1}.$$

An application of Slutsky's Theorem will thus yield our result, provided we can show that

$$n\tilde{X}^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}}-\tilde{X}\tilde{w}\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma_{F,\bar{w}}),\tag{69}$$

where

$$\Sigma_{F,\bar{w}}=\mathbb{E}\left[\frac{X_j^T\bar{w}(1-X_j^T\bar{w})}{\mu^T\bar{w}}\left(\frac{X_j}{X_j^T\mu}-\frac{\tilde{\Delta}\bar{w}}{2\mu^T\bar{w}}\right)\left(\frac{X_j}{X_j^T\mu}-\frac{\tilde{\Delta}\bar{w}}{2\mu^T\bar{w}}\right)^T\right].$$

To establish (69), we recall $t_v=\sum_{j=1}^{n}X_j^T\bar{w}=\mathbb{E}d_v$ and note that

$$\begin{aligned}
n\tilde{X}^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}}-\tilde{X}\tilde{w}\right)&=\frac{n\tilde{X}^TT^{-1/2}(\vec{a}-Xw)}{\sqrt{t_v}}+n\tilde{X}^T\left(\frac{D^{-1/2}}{\sqrt{d_v}}-\frac{T^{-1/2}}{\sqrt{t_v}}\right)X\bar{w}\\
&\quad+n\tilde{X}^T\left(\frac{D^{-1/2}}{\sqrt{d_v}}-\frac{T^{-1/2}}{\sqrt{t_v}}\right)(\vec{a}-Xw).
\end{aligned}$$

The last of these terms is $O(n^{-1/2}\log n)$ by a Hoeffding inequality followed by an application of Lemma 18, so that

$$n\tilde{X}^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}}-\tilde{X}\tilde{w}\right)=n\tilde{X}^TT^{-1/2}\frac{(\vec{a}-Xw)}{\sqrt{t_v}} + n\tilde{X}^T\left(\frac{D^{-1/2}}{\sqrt{d_v}}-\frac{T^{-1/2}}{\sqrt{t_v}}\right)X\bar{w}+O(n^{-1/2}\log n). \tag{70}$$

Multiplying numerator and denominator and applying Lemma 18, it holds for all $i\in[n]$

$$\begin{aligned}\frac{1}{\sqrt{d_i}}-\frac{1}{\sqrt{t_i}}&=\frac{t_i-d_i}{(\sqrt{d_i}+\sqrt{t_i})\sqrt{d_it_i}}=\frac{t_i-d_i}{2t_i^{3/2}}+(t_i-d_i)\frac{t_i(\sqrt{t_i}-\sqrt{d_i})+(t_i-d_i)\sqrt{t_i}}{2t_i^{3/2}(d_i\sqrt{t_i}+t_i\sqrt{d_i})}\\&=\frac{t_i-d_i}{2t_i^{3/2}}+O(n^{-3/2}\log n),\end{aligned}$$

and a similar result holds for the out-of-sample vertex, in that

$$\frac{1}{\sqrt{d_v}}-\frac{1}{\sqrt{t_v}}=\frac{t_v-d_v}{2t_v^{3/2}}+O(n^{-3/2}\log n).$$

Thus,

$$\begin{aligned}&\tilde{X}^T\left(\frac{D^{-1/2}}{\sqrt{d_v}}-\frac{T^{-1/2}}{\sqrt{t_v}}\right)X\bar{w}\\&=\tilde{X}^TT^{-1/2}\left(\frac{1}{\sqrt{d_v}}-\frac{1}{\sqrt{t_v}}\right)X\bar{w}+\tilde{X}^T\frac{(D^{-1/2}-T^{-1/2})X\bar{w}}{\sqrt{d_v}}\\&=\tilde{X}^TT^{-1/2}\frac{t_v-d_v}{2t_v^{3/2}}X\bar{w}+\frac{\tilde{X}^TT^{-3/2}(T-D)X\bar{w}}{2\sqrt{d_v}}+\sum_{j=1}^n\xi_jX_j^T\bar{w}\left(\frac{1}{\sqrt{t_j}}+\frac{1}{\sqrt{d_j}}\right)\frac{X_j}{\sqrt{t_j}}\end{aligned}$$

where $\xi_j\in\mathbb{R}, j=1,2,\ldots,n$ satisfy $\xi_j=O(n^{-3/2}\log n)$. Using Lemma 18, this last sum is itself $O(n^{-3/2}\log n)$, so that

$$n\tilde{X}^T\left(\frac{D^{-1/2}}{\sqrt{d_v}}-\frac{T^{-1/2}}{\sqrt{t_v}}\right)X\bar{w}=n\tilde{X}^TT^{-1/2}\frac{t_v-d_v}{2t_v^{3/2}}X\bar{w} + n\tilde{X}^T\frac{T^{-3/2}(T-D)X\bar{w}}{2\sqrt{d_v}}+O(n^{-1/2}\log n).$$

Plugging this into Equation (70),

$$n\tilde{X}^T\left(\frac{D^{-1/2}\vec{a}}{\sqrt{d_v}}-\tilde{X}\tilde{w}\right)=n\tilde{X}^TT^{-1/2}\frac{(\vec{a}-Xw)}{\sqrt{t_v}}+n\tilde{X}^TT^{-1/2}\frac{t_v-d_v}{2t_v^{3/2}}X\bar{w} + n\tilde{X}^T\frac{T^{-3/2}(T-D)X\bar{w}}{2\sqrt{d_v}}+O(n^{-1/2}\log n).$$

To complete our proof, it will suffice to show the following two facts:

$$n\tilde{X}^T T^{-1/2}\left(\frac{(\vec{a}-X\bar{w})}{\sqrt{t_v}}+\frac{t_v-d_v}{2t_v^{3/2}}X\bar{w}\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma_{F,\bar{w}}) \tag{71}$$

$$n\tilde{X}^T\frac{T^{-3/2}(T-D)X\bar{w}}{2\sqrt{d_v}}\xrightarrow{P}0 \tag{72}$$

To see the latter of these two points, observe that by our definitions of $d_i=\sum_{j=1}^n A_{i,j}$ and $t_i=\sum_{j=1}^n X_j^T X_i$,

$$\begin{aligned}n\tilde{X}^T\frac{T^{-3/2}(T-D)X\bar{w}}{2\sqrt{d_v}}&=\frac{n}{2\sqrt{d_v}}\sum_{i=1}^n\frac{(t_i-d_i)}{t_i^2}X_i^T\bar{w}X_i\\&=\frac{n}{2\sqrt{d_v}}\sum_{i=1}^n\frac{X_i^TX_i}{t_i^2}X_i^T\bar{w}X_i+\frac{n}{2\sqrt{d_v}}\sum_{1\le i<j\le n}(X_j^TX_i-A_{i,j})\left(\frac{X_i^T\bar{w}}{t_i^2}X_i+\frac{X_j^T\bar{w}}{t_j^2}X_j\right).\end{aligned}$$

The former of these two sums is $O(n^{-1/2})$ by an application of Lemma 18 and using the fact that $X_i\in\operatorname{supp}F$ are bounded. The latter of these two sums is, conditioned on $\{X_i\}_{i=1}^n$, a sum of independent 0-mean random variables, with $\|t_j^{-2}(X_j^TX_i-A_{i,j})X_j^T\bar{w}X_j\|\in[-Ct_j^{-2},Ct_j^{-2}]$ for all $j\in[n]$. Thus,

$$\Pr\left[\left|\sum_{1\le i<j\le n}t_j^{-2}(X_j^TX_i-A_{i,j})X_j^T\bar{w}X_j\right|\ge s\right]\le 2\exp\left\{\frac{-Cs^2}{\sum_{i<j}t_j^{-4}}\right\}.$$

Let $E_n=\{t_j\ge Cn:j=1,2,\dots,n\}$ denote the high-probability event of Lemma 18, for which we have $\Pr[E_n^c]\le Cn^{-2}$ for all suitably large n. Taking $s=Cn^{-1}\log^{1/2}n$ for suitably large $C>0$, letting $\mathbb{P}_{E_n}$ denote conditional probability $\Pr[\cdot\mid E_n]$,

$$\mathbb{P}_{E_n}\left[\left|\sum_{1\le i<j\le n}t_j^{-2}(X_j^TX_i-A_{i,j})X_j^T\bar{w}X_j\right|\ge Cn^{-1}\log^{1/2}n\right]\le Cn^{-2}.$$

Thus,

$$\begin{aligned}&\Pr\left[\left|\sum_{1\le i<j\le n}t_j^{-2}(X_j^TX_i-A_{i,j})X_j^T\bar{w}X_j\right|\ge Cn^{-1}\log^{1/2}n\right]\\&\le\mathbb{P}_{E_n}\left[\left|\sum_{1\le i<j\le n}t_j^{-2}(X_j^TX_i-A_{i,j})X_j^T\bar{w}X_j\right|\ge Cn^{-1}\log^{1/2}n\right]+\Pr[E_n^c]\\&\le Cn^{-2},\end{aligned}$$

and we conclude that, bounding $d_v^{-1/2}=O(n^{-1/2}\log^{1/2}n)$ by Lemma 18,

$$n\tilde{X}^T\frac{T^{-3/2}(T-D)X\bar{w}}{2\sqrt{d_v}}=O(n^{-1/2}\log^{1/2}n),$$

which establishes (72).

It remains only to prove Equation (71). Let $m_i = nX_i^T\mu$ for $i \in [n]$ and define the diagonal matrix

$$M = \mathrm{diag}(m_1, m_2, \ldots, m_n) \in \mathbb{R}^{n\times n}.$$

The argument in Lemma 18 allows us to bound $|t_v^{-1/2} - (n\mu^T\bar{w})^{-1/2}|$, so an argument similar to that above, wherein we apply Hoeffding's inequality followed by Lemma 18, implies

$$n\left(\frac{1}{\sqrt{t_v}} - \frac{1}{\sqrt{n\mu^T\bar{w}}}\right)\tilde{X}^T T^{-1/2}(\vec{a} - X\bar{w}) = O(n^{-1/2}\log n).$$

Lemma 18 also bounds $\max_i |t_i^{-1/2} - m_i^{-1/2}|$, whence

$$\frac{n\tilde{X}^T(T^{-1/2} - M^{-1/2})(\vec{a} - X\bar{w})}{\sqrt{n\mu^T\bar{w}}} = O(n^{-1/2}\log n).$$

The same Hoeffding-style argument once again yields, recalling that $\tilde{X} = T^{-1/2}X$,

$$\frac{nX^T(T^{-1/2} - M^{-1/2})M^{-1/2}(\vec{a} - X\bar{w})}{\sqrt{n\mu^T\bar{w}}} = O(n^{-1/2}\log n).$$

Combining the above three displays, the first term in the quantity of interest in Equation (71) is

$$\frac{n\tilde{X}^T T^{-1/2}(\vec{a} - X\bar{w})}{\sqrt{t_v}} = \frac{nX^T M^{-1}(\vec{a} - X\bar{w})}{\sqrt{n\mu^T\bar{w}}} + \tilde{O}(n^{-1/2}). \tag{73}$$

Turning to the second term on the left-hand side of Equation (71), rearranging terms and recalling the definition of $\tilde{\Delta} = \mathbb{E}X_1X_1^T/(X_1^T\mu)$,

$$\frac{n\tilde{X}^T T^{-1/2}(t_v - d_v)X\bar{w}}{2t_v^{3/2}} = \frac{n(t_v - d_v)\tilde{X}^T\tilde{X}\bar{w}}{2(n\mu^T\bar{w})^{3/2}} + \tilde{O}(n^{-1/2}) = \frac{n(t_v - d_v)\tilde{\Delta}\bar{w}}{2(n\mu^T\bar{w})^{3/2}} + \tilde{O}(n^{-1/2}),$$

where the first equality follows from Lemma 18 and the second equality follows from using (multivariate) Hoeffding's inequality to bound

$$\|\tilde{X}^T\tilde{X} - \tilde{\Delta}\| = \left\|\sum_{i=1}^n \frac{X_iX_i^T}{X_i^T\mu} - \tilde{\Delta}\right\| = O(n^{-1/2}\log^{1/2} n).$$

Thus, combining with Equation (73), the quantity on the left-hand side of Equation (71) is

$$n\tilde{X}^T T^{-1/2}\left(\frac{(\vec{a} - X\bar{w})}{\sqrt{t_v}} + \frac{t_v - d_v}{2t_v^{3/2}}X\bar{w}\right) = \frac{nX^T M^{-1}(\vec{a} - X\bar{w})}{\sqrt{n\mu^T\bar{w}}} + \frac{n(t_v - d_v)\tilde{\Delta}\bar{w}}{2(n\mu^T\bar{w})^{3/2}} + O(n^{-1/2}\log^{1/2} n).$$

Rearranging, and recalling $m_i = nX_i^T\mu$, $t_v = \sum_{j=1}^n X_j^T\bar{w}$ and $d_v = \sum_{j=1}^n a_j$,

$$\begin{aligned} n\tilde{X}^T T^{-1/2}\left(\frac{(\vec{a} - X\bar{w})}{\sqrt{t_v}} + \frac{t_v - d_v}{2t_v^{3/2}}X\bar{w}\right) &= n\sum_{j=1}^n \frac{a_j - X_j^T\bar{w}}{\sqrt{n\mu^T\bar{w}}}\left(\frac{X_j}{nX_j^T\mu} - \frac{\tilde{\Delta}\bar{w}}{2n\mu^T\bar{w}}\right) + O(n^{-1/2}\log^{1/2} n) \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^n \frac{(a_j - X_j^T\bar{w})}{\sqrt{\mu^T\bar{w}}}\left(\frac{X_j}{X_j^T\mu} - \frac{\tilde{\Delta}\bar{w}}{2\mu^T\bar{w}}\right) + O(n^{-1/2}\log^{1/2} n). \end{aligned}$$

Observe that this is a sum of n independent mean-zero random variables, so that by the multivariate CLT and Slutsky's Theorem,

$$n\tilde{X}^T T^{-1/2}\left(\frac{(\vec{a}-X\bar{w})}{\sqrt{t_v}}+\frac{t_v-d_v}{2t_v^{3/2}}X\bar{w}\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma_{F,\bar{w}}),$$

where

$$\Sigma_{F,\bar{w}}=\mathbb{E}\left[\frac{X_j^T\bar{w}(1-X_j^T\bar{w})}{\mu^T\bar{w}}\left(\frac{X_j}{X_j^T\mu}-\frac{\tilde{\Delta}\bar{w}}{2\mu^T\bar{w}}\right)\left(\frac{X_j}{X_j^T\mu}-\frac{\tilde{\Delta}\bar{w}}{2\mu^T\bar{w}}\right)^T\right],$$

completing the proof. ■

References

A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola. Distributed large-scale natural graph factorization. In *Proceedings of the 22nd International World Wide Web Conference*, pages 37–48, 2013.

A. Athreya, V. Lyzinski, D. J. Marchette, C. E. Priebe, D. L. Sussman, and M. Tang. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78:1–18, 2016.

A. Athreya, D. E. Fishkind, K. Levin, V. Lyzinski, Y. Park, Y. Qin, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold Regularization: A Geometric Framework for Learning from Examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Y. Bengio, J. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for LLE, ISOMAP, MDS, eigenmaps, and spectral clustering. In *Advances in Neural Information Processing Systems 16*, 2004.

P. Billingsley. *Probability and Measure*. Wiley, 1995.

I. Borg and P. J. F. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

J. Cape, M. Tang, and C. E. Priebe. On spectral embedding performance and elucidating network structure in stochastic block model graphs. *Network Science*, 7(3):269–291, 2019.

F. Chung. *Spectral Graph Theory*. Number 92 in Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics. American Mathematical Society, 1997.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30, 2006.

S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85(4633), 2000.

J. Fan, D. Wang, K. Wang, and Z. Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47:3009–3031, 2019.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2012.

J. C. Gower and G. B. Dijksterhuis. *Procrustes Problems*. Number 30 in Oxford Statistical Science Series. Oxford University Press, 2004.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

P. W. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

A. Jansen, G. Sell, and V. Lyzinski. Scalable out-of-sample extension of graph embeddings using deep neural networks. *Pattern Recognition Letters*, 94(15):1–6, 2017.

L. G. S. Jeub, P. Balachandran, M. A. Porter, P. J. Mucha, and M. W. Mahoney. Think locally, act locally: The detection of small, medium-sized, and large communities in large networks. *Physical Review E*, 91(012821), 2015.

D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M¿ Bogu ná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

K. Levin, A. Jansen, and B. Van Durme. Segmental acoustic indexing for zero resource keyword search. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

K. Levin, A. Athreya, M. Tang, V. Lyzinski, and C. E. Priebe. A central limit theorem for an omnibus embedding of random dot product graphs. *arXiv:1705.09355*, 2017.

K. Levin, F. Roosta-Khorasani, M. W. Mahoney, and C. E. Priebe. Out-of-sample extension of graph adjacency spectral embedding. In *Proceedings of ICML*, 2018.

K. Levin, A. Lodhia, and E. Levina. Recovering low-rank structure from multiple networks with unknown edge distributions. *arXiv:1906.07265*, 2019.

W. Liu, J. He, and S.-F. Chang. Large graph construction for scalable semi-supervised learning. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

V. Lyzinski, D. L. Sussman, M. Tang, A. Athreya, and C. E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922, 2014.

V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions in Network Science and Engineering*, 4(1):13–26, 2017.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2002.

R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv:0911.0600*, 2010.

A. M. Quispe, C. Petitjean, and L. Heutte. Extreme learning machine for out-of-sample extension in laplacian eigenmaps. *Pattern Recognition Letters*, 74:68–73, 2016.

P. Rubin-Delanchy, C. E. Priebe, M. Tang, and J. Cape. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv 1709.05506*, 2017.

B. Srinivasan and B. Ribero. On the equivalence between positional node embeddings and structural graph representations. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

M. Tang and C. E. Priebe. Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415, 2018.

M. Tang, Y. Park, and C. E. Priebe. Out-of-sample extension of latent position graphs. *arXiv:1305.4893*, 2013a.

M. Tang, D. L. Sussman, and C. E. Priebe. Universally consistent vertex classification for latent position graphs. *The Annals of Statistics*, 31:1406–1430, 2013b.

M. Tang, J. Cape, and C. E. Priebe. Asymptotically efficient estimators for stochastic blockmodels: the naive MLE, the rank-constrained MLE, and the spectral. *arXiv:1710.10936*, 2017.

R. Tang, M. Ketcha, A. Badea, E. D. Calabrese, D. S. Margulies, J. T. Vogelstein, C. E. Priebe, and D. L. Sussman. Connectome smoothing via low-rank approximations. *IEEE Transactions on Medical Imaging*, 38(6):1446–1456, 2019.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4): 401–419, 1952.

J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

M. W. Trosset and C. E. Priebe. The out-of-sample problem for classical multidimensional scaling. *Computational Statistics and Data Analysis*, 52(10):4635–4642, 2008.

N. K. Vishnoi. Lx = b. *Foundations and Trends in Theoretical Computer Science*, 8(1–2): 1–141, 2013.

Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proc. IEEE International Conference on Computer Vision*, pages 975–982, 1999.

L. Xiao, S. Boyd, and S. Lall. A scheme for robust distributed sensor fusion based on average consensus. In *Fourth International Symposium on Information Processing in Sensor Networks*, 2005.

F. Xie and Y. Xu. Optimal Bayesian estimation for random dot product graphs. *Biometrika*, 107(4):875–889, 2020.

S. Young and E. Scheinerman. Random dot product graph models for social networks. In *Proceedings of the 5th International Conference on Algorithms and Models for the Web-graph*, pages 138–149, 2007.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102:315–323, 2015.

D. Zheng, D. Mhembere, R. Burns, J. Vogelstein C. E. Priebe, and A. S. Szalay. FlashGraph: Processing billion-node graphs on an array of commodity SSDs. In *13th USENIX Conference on File and Storage Technologies (FAST 15)*, pages 45–58, 2015.

M. Zhu and A. Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis*, 51:918–930, 2006.