Universal Latent Space Model Fitting for Large Networks with Edge Covariates

Zhuang Ma*, Zongming Ma†, and Hongsong Yuan‡

August 26, 2018

Abstract

Latent space models are effective tools for statistical modeling and visualization of network data. Due to their close connection to generalized linear models, it is also natural to incorporate covariate information in them. The current paper presents two universal fitting algorithms for networks with edge covariates: one based on nuclear norm penalization and the other based on projected gradient descent. Both algorithms are motivated by maximizing the likelihood function for an existing class of inner-product models, and we establish their statistical rates of convergence for these models. In addition, the theory informs us that both methods work simultaneously for a wide range of different latent space models that allow latent positions to affect edge formation in flexible ways, such as distance models. Furthermore, the effectiveness of the methods is demonstrated on a number of real world network datasets for different statistical tasks, including community detection with and without edge covariates, and network assisted learning.

Keywords: community detection, network with covariates, non-convex optimization, projected gradient descent.

1 Introduction

Network is a prevalent form of data for quantitative and qualitative analysis in a number of fields, including but not limited to sociology, computer science, neuroscience, etc. Moreover, due to advances in science and technology, the sizes of the networks we encounter are ever increasing. Therefore, to explore, to visualize and to utilize the information in large networks poses significant challenges to Statistics. Unlike traditional datasets in which a number of features are recorded for each subject, network datasets provide information on the relation among all subjects under study, sometimes together with additional features. In this paper, we focus on networks in which additional features might be observed for each node pair.

An efficient way to extract key information from network data is to fit appropriate statistical models on them. To date, there have been a collection of network models proposed by researchers in various fields. These models aim to catch different aspects of network characteristics, and Goldenberg et al. [29] provides a comprehensive overview. An important class of network models

*Department of Statistics, University of Pennsylvania.

†Department of Statistics, University of Pennsylvania. Email: zongming@wharton.upenn.edu.

‡School of Information Management and Engineering, Shanghai University of Finance and Economics.

are *latent space models* which was proposed in an influential paper by Hoff et al. [34]. Suppose there are n nodes in the observed network. The key idea underlying latent space modeling is that each node i can be represented by a vector z_i in some low dimensional Euclidean space (or some other metric space of choice, see e.g. [41, 6] for latent spaces with negative curvature) that is sometimes called the social space, and nodes that are "close" in the social space are more likely to be connected. Hoff et al. [34] considered two types of latent space models: distance models and projection models. In both cases, the latent vectors $\{z_i\}_{i=1}^n$ were treated as fixed effects. Later, a series of papers [31, 30, 33, 43] generalized the original proposal in [34] for better modeling of other characteristics of social networks, such as clustering, degree heterogeneity, etc. In these generalizations, the z_i's were treated as random effects generated from certain multivariate Gaussian mixtures. Moreover, model fitting and inference in these models has been carried out via Markov Chain Monte Carlo, and it is difficult to scale these methodologies to handle large networks [29]. In addition, one needs to use different likelihood function based on choice of model and there is little understanding of the quality of fitting when the model is mis-specified. Albeit these disadvantages, latent space models are attractive due to their friendliness to interpretation and visualization.

For concreteness, assume that we observe an undirected network represented by a symmetric adjacency matrix A on n nodes with $A_{ij} = A_{ji} = 1$ if nodes i and j are connected and zero otherwise. In addition, we may also observe a symmetric pairwise covariate matrix X which measures certain characteristics of node pairs. We do not allow self-loop and so we set the diagonal elements of the matrices A and X to be zeros. The covariate X_{ij} can be binary, such as an indicator of whether nodes i and j share some common attribute (e.g. gender, location, etc) or it can be continuous, such as a distance/similarity measure (e.g. difference in age, income, etc). It is relatively straightforward to generalize the methods and theory in this paper to multiple covariates.

1.1 Main contributions

The main contributions of the present paper are the following.

1. We first consider an existing class of latent space models, called inner-product models [31, 32], and design two new fitting algorithms for this class. Let the observed n-by-n adjacency matrix and covariate matrix be A and X, respectively. The inner-product model assumes that for any $i < j$,

$$\begin{aligned} A_{ij} = A_{ji} &\overset{ind.}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with} \\ \text{logit}(P_{ij}) = \Theta_{ij} &= \alpha_i + \alpha_j + \beta X_{ij} + z_i^\top z_j, \end{aligned} \tag{1}$$

   where for any $x \in (0,1)$, $\text{logit}(x) = \log[x/(1-x)]$. Here, α_i, $1 \leqslant i \leqslant n$, are parameters modeling degree heterogeneity. The parameter β is the coefficient for the observed covariate, and $z_i^\top z_j$ is the inner-product between the latent vectors. From a matrix estimation viewpoint, the matrix $G = (G_{ij}) = (z_i^\top z_j)$ is of rank at most k that can be much smaller than n. Motivated by recent advances in low rank matrix estimation, we design two algorithms for fitting (1). One algorithm is based on lifting and nuclear norm penalization of the negative log-likelihood function. The other is based on directly optimizing the negative log-likelihood function via projected gradient descent. The methods can be used to fit these models on networks with thousands of nodes easily on any reasonable personal computer and has the potential to scale to even larger networks. For both algorithms, we establish high probability error bounds for inner-product models. The connection between model (1) and the associated

algorithms and other related work in the literature will be discussed immdediately in next subsection.

2. More importantly, we further show that these two fitting algorithms are "universal" in the sense that they can work simultaneously for a wide range of latent space models beyond the inner-product model class. For example, they work for the distance model and the Gaussian kernel model in which the inner-product term $z_i^\top z_j$ in (1) is replaced with $-\|z_i - z_j\|$ and $c\exp(-\|z_i - z_j\|^2/\sigma^2)$, respectively. Thus, the class of inner-product models is flexible and can be used to approximate many other latent space models of interest. In addition, the associated algorithms can be applied to networks generated from a wide range of mis-specified models and still yield reasonable results. The key mathematical insight that enables such universality is introduced in Section 2 as the Schoenberg Condition (7).

3. We demonstrate the effectiveness of the model and algorithms on real data examples. In particular, we fit inner-product models by the proposed algorithms on five different real network datasets for several different tasks, including visualization, clustering and network-assisted classification. On three popular benchmark datasets for testing community detection on networks, a simple k-means clustering on the estimated latent vectors obtained by our algorithm yields as good result on one dataset and better results on the other two when compared with four state-of-the-art methods. The same "model fitting followed by k-means clustering" approach also yields nice clustering of nodes on a social network with edge covariates. Due to the nature of latent space models, for all datasets on which we fit the model, we obtain natural visualizations of the networks by plotting latent positions. Furthermore, we illustrate how network information can be incorporated in traditional learning problems using a document classification example.

A Matlab implementation of the methods in the present paper is available upon request.

1.2 Other related works and issues

The current form of the inner-product model (1) has previously appeared in Hoff [31] and Hoff [32], though the parameters were modeled as random effects rather than fixed values, and Bayesian approaches were proposed for estimating variance parameters of the random effects. Hoff [33] proposed a latent eigenmodel which has a probit link function as opposed to the logistic link function in the present paper. As in [31] and [32], parameters were modeled as random effects and model fitting was through Bayesian methods. It was shown that the eigenmodel weakly generalizes the distance model in the sense that the order of the entries in the latent component can be preserved. This is complementary to our results which aim to approximate the latent component directly in some matrix norm. An advantage of the eigenmodel is its ability to generalize the latent class model, whereas the inner-product model (1) and the more general model we shall consider generalize a subset of latent class models due to the constraint that the latent component (after centering) is positive semi-definite. We shall return to this point later in Section 7. Young and Scheinerman [63] proposed a random dot product model, which can be viewed as an inner-product model with identity link function. The authors studied a number of properties of the model, such as diameter, clustering and degree distribution. See also [56] for some statistical theory for this model. Tang et al. [57] studied properties of the leading eigenvectors of the adjacency matrices of latent positions graphs (together with its implication on classification in such models) where

the connection probability of two nodes is the value that some universal kernel [49] takes on the associated latent positions and hence generalizes the random dot product model. This work is close in spirit to the present paper. However, there are several important differences. First, the focus here is model fitting/parameter estimation as opposed to classification in [57]. In addition, any universal kernel considered in [57] satisfies the Schoenberg condition (7) and thus is covered by the methods and theory of the present paper, and so we cover a broader range of models that inner-product models can approximate. This is also partially due to the different inference goals. Furthermore, we allow the presence of observed covariates while [57] did not.

When fitting a network model, we are essentially modeling and estimating the edge probability matrix. From this viewpoint, the present paper is related to the literautre on graphon estimation and edge probability matrix estimation for block models. See, for instance, [7, 4, 61, 25, 39, 27] and the references therein. However, the block models have stronger structural assumptions than the latent space models we are going to investigate.

The algorithmic and theoretical aspects of the paper is also closely connected to the line of research on low rank matrix estimation, which plays an important role in many applications such as phase retrieval [14, 15] and matrix completion [13, 37, 38, 12, 40]. Indeed, the idea of nuclear norm penalization has originated from matrix completion for both general entries [13] and binary entries [22]. In particular, our convex approach can be viewed as a Lagrangian form of the proposal in [22] when there is no covariate and the matrix is fully observed. We have nonetheless decided to spell out details on both method and theory for the convex approach because the matrix completion literature typically does not take into account the potential presence of observed covariates. On the other hand, the idea of directly optimizing a non-convex objective function involving a low rank matrix has been studied recently in a series of papers. See, for instance, [46, 11, 55, 60, 18, 66, 28] and the references therein. Among these papers, the one that is the most related to the projected gradient descent algorithm we are to propose and analyze is [18] which focused on estimating a positive semi-definite matrix of exact low rank in a collection of interesting problems. Another recent and related work [62] has appeared after the initial posting of the present manuscript. However, we will obtain tighter error bounds for latent space models and we will go beyond the exact low rank scenario.

From a link prediction viewpoint, it is natural to incorporate edge covariates. With appropriate regularization to guard against overfitting, such extra information contained in edge covariates can usually improve prediction performance. On the other hand, one may want to conduct community detection after fitting latent space models. When this is the case, incorporation of edge covariates becomes a more subtle issue. If one only uses edge covariates that are independent of the community structure, then including them when fitting the model should help the estimation of latent variables and hence community detection. However, when one incorporates edge covariates that are highly dependent on community assignment, their inclusion may worsen the performance of community detection by latent variables alone, and it is more reasonable to perform community detection using both observed and latent variables. Since this is highly case dependent, we shall not attempt a general treatment along this direction.

1.3 Organization

After a brief introduction of standard notation used throughout the paper, the rest of the paper is organized as follows. Section 2 introduces both inner-product models and a broader class of latent

space models on which our fitting methods work. The two fitting methods are described in detail in Section 3, followed by their theoretical guarantees in Section 4 under both inner-product models and the general class. The theoretical results are further corroborated by simulated examples in Section 5. Section 6 demonstrates the competitive performance of the modeling approach and fitting methods on five different real network datasets. We discuss interesting related problems in Section 7 and present proofs of the main results in Section 8. Technical details justifying the initialization methods for the project gradient descent approach are deferred to the appendix. Furthermore, the appendix also discusses some method for dealing with multiple edge covariates.

Notation For $A = (A_{ij}) \in \mathbb{R}^{n\times n}$, $\mathrm{Tr}(A) = \sum_{i=1}^n A_{ii}$ stands for its trace. For $X, Y \in \mathbb{R}^{m\times n}$, $\langle X, Y\rangle = \mathrm{Tr}(X^\top Y)$ defines an inner product between them. If $m \geqslant n$, for any matrix X with singular value decomposition $X = \sum_{i=1}^n \sigma_i u_i v_i^\top$, $\|X\|_* = \sum_{i=1}^n \sigma_i$, $\|X\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^n \sigma_i^2}$ and $\|X\|_{\mathrm{op}} = \max_{i=1}^n \sigma_i$ stand for the nuclear norm, the Frobenius norm and the operator norm of the matrix, respectively. Moreover, X_{i*} and X_{*j} denote the i-th row and j-th column of X, and for any function f, $f(X)$ is the shorthand for applying $f(\cdot)$ element-wisely to X, that is $f(X) \in \mathbb{R}^{m\times n}$ and $[f(X)]_{ij} = f(X_{ij})$. Let $\mathbb{S}_+^n$ be the set of all $n \times n$ positive semidefinite matrices and $O(m, n)$ be the set of all $m \times n$ orthonormal matrices. For any convex set $\mathcal{C}$, $P_{\mathcal{C}}(\cdot)$ is the projection onto the $\mathcal{C}$.

2 Latent space models

In this section, we first give a detailed introduction of the inner-product model (1) and conditions for its identifiability. In addition, we introduce a more general class of latent space models that includes the inner-product model as a special case. The methods we propose later will be motivated by the inner-product model and can also be applied to the more general class.

2.1 Inner-product models

Recall the inner-product model defined in (1), i.e., for any observed A and X and any $i < j$,

$$A_{ij} = A_{ji} \overset{ind.}{\sim} \mathrm{Bernoulli}(P_{ij}), \quad \text{with} \quad \mathrm{logit}(P_{ij}) = \Theta_{ij} = \alpha_i + \alpha_j + \beta X_{ij} + z_i^\top z_j.$$

Fixing all other parameters, if we increase α_i, then node i has higher chances of connecting with other nodes. Therefore, the α_i's model degree heterogeneity of nodes and we call them degree heterogeneity parameters. Next, the regression coefficient β moderates the contribution of covariate to edge formation. For instance, if X_{ij} indicates whether nodes i and j share some common attribute such as gender, then a positive β value implies that nodes that share common attribute are more likely to connect. Such a phenomenon is called *homophily* in the social network literaute. Last but not least, the latent variables $\{z_i\}_{i=1}^n$ enter the model through their inner-product $z_i^\top z_j$, and hence is the name of the model. We impose no additional structural/distributional assumptions on the latent variables for the sake of modeling flexibility.

We note that model (1) also allows the latent variables to enter the second equation in the form of $g(z_i, z_j) = -\frac{1}{2}\|z_i - z_j\|^2$. To see this, note that $g(z_i, z_j) = -\frac{1}{2}\|z_i\|^2 - \frac{1}{2}\|z_j\|^2 + z_i^\top z_j$, and we may re-parameterize by setting $\tilde{\alpha}_i = \alpha_i - \frac{1}{2}\|z_i\|^2$ for all i. Then we have

$$\Theta_{ij} = \alpha_i + \alpha_j + \beta X_{ij} - \frac{1}{2}\|z_i - z_j\|^2 = \tilde{\alpha}_i + \tilde{\alpha}_j + \beta X_{ij} + z_i^\top z_j.$$

An important implication of this observation is that the function $g(z_i, z_j) = -\frac{1}{2}\|z_i - z_j\|^2$ directly models *transitivity*, i.e., nodes with common neighbors are more likely to connect since their latent variables are more likely to be close to each other in the latent space. In view of the foregoing discussion, the inner-product model (1) also enjoys this nice modeling capacity.

In matrix form, we have

$$\Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + G \tag{2}$$

where 1_n is the all one vector in $\mathbb{R}^n$ and $G = ZZ^\top$ with $Z = (z_1, \cdots, z_n)^\top \in \mathbb{R}^{n\times k}$. Since there is no self-edge and Θ is symmetric, only the upper diagonal elements of Θ are well defined, which we denote by Θ^u. Nonetheless we define the diagonal element of Θ as in (2) since it is inconsequential. To ensure identifiability of model parameters in (1), we assume the latent variables are centered, that is

$$JZ = Z \quad \text{where} \quad J = I_n - \frac{1}{n} 1_n 1_n{}^\top. \tag{3}$$

Note that this condition uniquely identifies Z up to a common orthogonal transformation of its rows while $G = ZZ^\top$ is now directly identifiable.

2.2 A more general class and the Schoenberg condition

Model (1) is a special case of a more general class of latent space models, which can be defined by

$$\begin{aligned} A_{ij} = A_{ji} &\overset{ind.}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with} \\ \text{logit}(P_{ij}) = \Theta_{ij} &= \tilde{\alpha}_i + \tilde{\alpha}_j + \beta X_{ij} + \ell(z_i, z_j) \end{aligned} \tag{4}$$

where $\ell(\cdot, \cdot)$ is a smooth symmetric function on $\mathbb{R}^k \times \mathbb{R}^k$. We shall impose an additional constraint on ℓ following the discussion below. In matrix form, for $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)'$ and $L = (\ell(z_i, z_j))$, we can write

$$\Theta = \tilde{\alpha} 1_n{}^\top + 1_n \tilde{\alpha}^\top + \beta X + L.$$

To better connect with (2), let

$$G = JLJ, \quad \text{and} \quad \alpha 1_n{}^\top + 1_n \alpha^\top = \tilde{\alpha} 1_n{}^\top + 1_n \tilde{\alpha}^\top + L - JLJ. \tag{5}$$

Note that the second equality in the last display holds since the expression on its righthand side is symmetric and of rank at most two. Then we can rewrite the second last display as

$$\Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + G \tag{6}$$

which reduces to (2) and G satisfies $JG = G$. Our additional constraint on ℓ is the following Schoenberg condition:

$$\begin{aligned} &\text{For any positive integer } n \geqslant 2 \text{ and any } z_1, \ldots, z_n \in \mathbb{R}^k, \\ &G = JLJ \text{ is positive semi-definite for } L = (\ell(z_i, z_j)) \text{ and } J = I_n - \tfrac{1}{n} 1_n 1_n{}^\top. \end{aligned} \tag{7}$$

Condition (7) may seem abstract, while the following lemma elucidates two important classes of symmetric functions for which it is satisfied.

Lemma 2.1. *Condition* (7) *is satisfied in the following cases:*

1. *ℓ is a positive semi-definite kernel function on $\mathbb{R}^k \times \mathbb{R}^k$;*

2. *$\ell(x, y) = -\|x-y\|_p^q$ for some $p \in (0, 2]$ and $q \in (0, p]$ where $\|\cdot\|_p$ is the p-norm (or p-seminorm when $p < 1$) on $\mathbb{R}^k$.*

The first claim of Lemma 2.1 is a direct consequence of the definition of positive semi-definite kernel function which ensures that the matrix L itself is positive semi-definite and so is $G = JLJ$ since J is also positive semi-definite. The second claim is a direct consequence of the Hilbert space embedding result by Schoenberg [53, 54]. See, for instance, Theorems 1 and 2 of [53].

3 Two model fitting methods

This section presents two methods for fitting models (1) and (4)–(7). Both methods are motivated by minimizing the negative log-likelihood function of the inner-product model, and can be regarded as pseudo-likelihood approaches for more general models. From a methodological viewpoint, a key advantage of these methods, in particular the projected gradient descent method, is the scalability to networks of large sizes.

3.1 A convex approach via penalized MLE

In either the inner-product model or the general model we suppose the parameter matrix Θ in (2) or (6) satisfies

$$-M_1 \leqslant \Theta_{ij} \leqslant -M_2 \;\text{ for }\; 1 \leqslant i \neq j \leqslant n, \quad \text{and} \quad |\Theta_{ii}| \leqslant M_1 \;\text{ for }\; 1 \leqslant i \leqslant n. \tag{8}$$

where $M_1 \geqslant M_2$ are non-negative. Then for any Θ satisfying (8), the corresponding edge probabilities satisfy

$$\frac{1}{2}e^{-M_1} \leqslant \frac{1}{1+e^{M_1}} \leqslant P_{ij} \leqslant \frac{1}{1+e^{M_2}} \leqslant e^{-M_2},\ 1 \leqslant i \neq j \leqslant n. \tag{9}$$

Thus M_1 controls the conditioning of the problem and M_2 controls the sparsity of the network.

Let $\sigma(x) = 1/(1+e^{-x})$ be the sigmoid function, i.e. the inverse of logit function, then for any $i \neq j$, $P_{ij} = \sigma(\Theta_{ij})$ and the log-likelihood function of the inner-product model (1) can be written as

$$\sum_{i<j} \Big\{ A_{ij} \log\Big(\sigma(\Theta_{ij})\Big) + (1-A_{ij}) \log\Big(1-\sigma(\Theta_{ij})\Big) \Big\} = \sum_{i<j} \Big\{ A_{ij}\Theta_{ij} + \log\Big(1-\sigma(\Theta_{ij})\Big) \Big\}.$$

Recall that $G = ZZ^\top$ in inner-product models. The MLE of Θ^u is the solution to the following rank constrained optimization problem:

$$\begin{aligned} \min_{\Theta^u,\alpha,\beta,G} \quad & -\sum_{i<j} \Big\{ A_{ij}\Theta_{ij} + \log\Big(1-\sigma(\Theta_{ij})\Big) \Big\}, \\ \text{subject to} \quad & \Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + G, \quad -M_1 \leqslant \Theta_{ij} \leqslant -M_2, \\ & GJ = G, \quad G \in \mathbb{S}_+^n, \quad \operatorname{rank}(G) \leqslant k. \end{aligned} \tag{10}$$

This optimization problem is non-convex and generally intractable. To overcome this difficulty, we consider a convex relaxation that replaces the rank constraint on G in (10) with a penalty term

on its nuclear norm. Since G is positive semi-definite, its nuclear norm equals its trace. Thus, our first model fitting scheme solves the following convex program:

$$\begin{aligned} \min_{\alpha,\beta,G} \quad & -\sum_{i,j}\Big\{A_{ij}\Theta_{ij} + \log\Big(1-\sigma(\Theta_{ij})\Big)\Big\} + \lambda_n \operatorname{Tr}(G) \\ \text{subject to} \quad & \Theta = \alpha 1_n{}^\top + 1_n\alpha^\top + \beta X + G, GJ = G, \;\; G \in \mathbb{S}_+^n, \quad -M_1 \leqslant \Theta_{ij} \leqslant -M_2. \end{aligned} \tag{11}$$

The convex model fitting method (11) is motivated by the nuclear norm penalization idea originated from the matrix completion literature. See, for instance, [13], [12], [40], [22] and the references therein. In particular, it can be viewed as a Lagrangian form of the proposal in [22] when there is no covariate and the matrix is fully observed. However, we have decided to make this proposal and study the theoretical properties as the existing literature, such as [22], does not take in consideration the potential presence of observed covariates. Furthermore, one can still solve (11) when the true underlying model is one of the general models introduced in Section 2.2. The appropriate choice of λ_n will be discussed in Section 4.

Remark 3.1. In addition to the introduction of the trace penalty, the first term in the objective function in (11) now sums over all (i, j) pairs. Due to symmetry, after scaling, the difference from the sum in (10) lies in the inclusion of all diagonal terms in Θ. This slight modification leads to neither theoretical consequence nor noticeable difference in practice. However, it allows easier implementation and simplifies the theoretical investigation. We would note that the constraint $-M_1 \leqslant \Theta_{ij} \leqslant -M_2$ is included partially for obtaining theoretical guarantees. In simulated examples reported in Section 5, we found that the convex program worked equally well without this constraint.

3.2 A non-convex approach via projected gradient descent

Although the foregoing convex relaxation method is conceptually neat, state-of-the-art algorithms to solve the nuclear (trace) norm minimization problem (11), such as iterative singular value thresholding, usually require computing a full singular value decomposition at every iteration, which can still be time consuming when fitting very large networks.

To further improve scalability of model fitting, we propose an efficient first order algorithm that directly tackles the following non-convex objective function:

$$\min_{Z,\alpha,\beta} \; g(Z,\alpha,\beta) = -\sum_{i,j}\Big\{A_{ij}\Theta_{ij} + \log\Big(1-\sigma(\Theta_{ij})\Big)\Big\} \quad \text{where } \Theta = \alpha 1_n{}^\top + 1_n\alpha^\top + \beta X + ZZ^\top. \tag{12}$$

The detailed description of the method is presented in Algorithm 1.

After initialization, Algorithm 1 iteratively updates the estimates for the three parameters, namely Z, α and β. In each iteration, for each parameter, the algorithm first descends along the gradient direction by a pre-specified step size. The descent step is then followed by an additional projection step which projects the updated estimates to pre-specified constraint sets. We propose to set the step sizes as

$$\eta_Z = \eta/\left\|Z^0\right\|_{\mathrm{op}}^2, \quad \eta_\alpha = \eta/(2n), \quad \text{and} \quad \eta_\beta = \eta/(2\left\|X\right\|_{\mathrm{F}}^2) \tag{13}$$

for some small numeric constant $\eta > 0$. To establish the desired theoretical guarantees, we make a specific choice of the constraint sets later in the statement of Theorem 4.2 and Theorem 4.4. In practice, one may simply set

$$Z^{t+1} = J\widetilde{Z}^{t+1}, \quad \alpha^{t+1} = \widetilde{\alpha}^{t+1}, \quad \text{and} \quad \beta^{t+1} = \widetilde{\beta}^{t+1}. \tag{14}$$

Algorithm 1 A projected gradient descent model fitting method.

Input: Adjacency matrix: A; covariate matrix: X; latent space dimension: $k \geqslant 1$; initial estimates: Z^0, α^0, β^0; step sizes: $\eta_Z, \eta_\alpha, \eta_\beta$; constraint sets: $\mathcal{C}_Z, \mathcal{C}_\alpha, \mathcal{C}_\beta$.

Output: $\widehat{Z} = Z^T$, $\widehat{\alpha} = \alpha^T$, $\widehat{\beta} = \beta^T$.

for $t = 0, 1, \cdots, T-1$ **do**

$\widetilde{Z}^{t+1} = Z^t - \eta_Z \nabla_Z g(Z, \alpha, \beta) = Z^t + 2\eta_Z \left(A - \sigma(\Theta^t)\right) Z^t$;

$\widetilde{\alpha}^{t+1} = \alpha^t - \eta_\alpha \nabla_\alpha g(Z, \alpha, \beta) = \alpha^t + 2\eta_\alpha (A - \sigma(\Theta^t)) 1_n$;

$\widetilde{\beta}^{t+1} = \beta^t - \eta_\beta \nabla_\beta g(Z, \alpha, \beta) = \beta^t + \eta_\beta \langle A - \sigma(\Theta^t), X \rangle$;

$Z^{t+1} = \mathcal{P}_{\mathcal{C}_Z}(\widetilde{Z}^{t+1})$, $\alpha^{t+1} = \mathcal{P}_{\mathcal{C}_\alpha}(\widetilde{\alpha}^{t+1})$, $\beta^{t+1} = \mathcal{P}_{\mathcal{C}_\beta}(\widetilde{\beta}^{t+1})$;

end for

Here and after, when there is no covariate, i.e. $X = 0$, we skip the update of β in each iteration.

For each iteration, the update on the latent part is performed in the space of Z (that is $\mathbb{R}^{n \times k}$) rather than the space of all $n \times n$ Gram matrices as was required in the convex approach. In this way, it reduces the computational cost per iteration from $O(n^3)$ to $O(n^2 k)$. Since we are most interested in cases where $k \ll n$, such a reduction leads to improved scalability of the non-convex approach to large networks. To implement this non-convex algorithm, we need to specify the latent space dimension k, which was not needed for the convex program (11). We defer the discussion on the data-driven choice of k to Section 7.

We note that Algorithm 1 is not guaranteed to find any global minimizer, or even any local minimizer, of the objective function (12). However, as we shall show later in Section 4, under appropriate conditions, the iterates generated by the algorithm will quickly enter a neighborhood of the true parameters $(Z_\star, \alpha_\star, \beta_\star)$ and any element in this neighborhood is statistically at least as good as the estimator obtained from the convex method (11). This approach has close connection to the investigation of various non-convex methods for other statistical and signal processing applications. See for instance [15], [18] and the references therein. Our theoretical analysis of the algorithm is going to provide some additional insight as we shall establish its high probability error bounds for both the exact and the approximate low rank scenarios. In the rest of this section, we discuss initialization of Algorithm 1.

3.2.1 Initialization

Appropriate initialization is the key to success for Algorithm 1. We now present two ways to initialize it which are theoretically justifiable under different circumstances.

Initialization by projected gradient descent in the lifted space The first initialization method is summarized in Algorithm 2, which is essentially running the projected gradient descent algorithm on the following regularized objective function for a small number of steps:

$$f(G, \alpha, \beta) = -\sum_{i,j} \{A_{ij}\Theta_{ij} + \log(1 - \sigma(\Theta_{ij}))\} + \lambda_n \operatorname{Tr}(G) + \frac{\gamma_n}{2}\left(\|G\|_{\mathrm{F}}^2 + 2\left\|\alpha 1_n{}^\top\right\|_{\mathrm{F}}^2 + \|X\beta\|_{\mathrm{F}}^2 \right). \quad (15)$$

Except for the third term, this is the same as the objective function in (11). However, the inclusion of the additional proximal term ensures that one obtains the desired initializers after a minimal number of steps.

Algorithm 2 Initialization of Algorithm 1 by Projected Gradient Descent

Input: Adjacency matrix: A; covariate matrix X; initial values: $G^0 = 0, \alpha^0 = 0, \beta^0 = 0$; step size: η; constraint set: $\mathcal{C}_G, \mathcal{C}_\alpha, \mathcal{C}_\beta$; regularization parameter: λ_n, γ_n; latent dimension: k; number of steps: T.
for $t = 1, 2, \cdots, T$ **do**
  $\widetilde{G}^{t+1} = G^t - \eta \nabla_G f(G, \alpha, \beta) = G^t + \eta(A - \sigma(\Theta^t) - \lambda_n I_n - \gamma_n G^t)$;
  $\widetilde{\alpha}^{t+1} = \alpha^t - \eta \nabla_\alpha f(G, \alpha, \beta)/n = \alpha^t + \eta((A - \sigma(\Theta^t))1_n/2n - \gamma_n \alpha^t)$;
  $\widetilde{\beta}^{t+1} = \beta^t - \eta \nabla_\beta f(G, \alpha, \beta)/\|X\|_{\mathrm{F}}^2 = \beta^t + \eta(\langle A - \sigma(\Theta^t), X \rangle/\|X\|_{\mathrm{F}}^2 - \gamma_n \beta^t)$;
  $G^{t+1} = \mathcal{P}_{\mathcal{C}_G}(\widetilde{G}^{t+1}),\ \alpha^{t+1} = \mathcal{P}_{\mathcal{C}_\alpha}(\widetilde{\alpha}^{t+1}),\ \beta^{t+1} = \mathcal{P}_{\mathcal{C}_\beta}(\widetilde{\beta}^{t+1})$;
end for
Set $Z^T = U_k D_k^{1/2}$ where $U_k D_k U_k^\top$ is the top-k eigen components of G^T;
Output: Z^T, α^T, β^T.

The appropriate choices of λ_n and γ_n will be spelled out in Theorem 4.5 and Corollary 4.1. The step size η in Algorithm 2 can be set at a small positive numeric constant, e.g. $\eta = 0.2$. The projection sets that lead to theoretical justification will be specified later in Theorem 4.5 while in practice, one may simply set $G^{t+1} = J\widetilde{G}^{t+1}$, $\alpha^{t+1} = \widetilde{\alpha}^{t+1}$, and $\beta^{t+1} = \widetilde{\beta}^{t+1}$.

Initialization by universal singular value thresholding Another way to construct the initialization is to first estimate the probability matrix P by universal singular value thresholding (USVT) proposed by [16] and then compute the initial estimates of α, Z, β heuristically by inverting the logit transform. The procedure is summarized as Algorithm 3.

Algorithm 3 Initialization of Algorithm 1 by Singular Value Thresholding

Input: Adjacency matrix: A; covariate matrix X; latent dimension k; threshold τ.
Let $\widetilde{P} = \sum_{\sigma_i \geqslant \tau} \sigma_i u_i v_i^\top$ where $\sum_{i=1}^n \sigma_i u_i v_i^\top$ is the SVD of A. Elementwisely project $\widetilde{P}$ to the interval $[\frac{1}{2}e^{-M_1}, \frac{1}{2}]$ to obtain $\widehat{P}$. Let $\widehat{\Theta} = \mathrm{logit}((\widehat{P} + \widehat{P}^\top)/2)$;
Let $\alpha^0, \beta^0 = \arg\min_{\alpha,\beta} \|\widehat{\Theta} - \left(\alpha 1_n{}^\top + 1_n \alpha^\top + \beta X\right)\|_{\mathrm{F}}^2$;
Let $\widehat{G} = \mathcal{P}_{\mathbb{S}_+^n}(R)$ where $R = J(\widehat{\Theta} - (\alpha^0 1_n{}^\top + 1_n(\alpha^0)^\top + \beta^0 X))J$;
Set $Z^0 = U_k D_k^{1/2}$ where $U_k D_k U_k^\top$ is the top-k singular value components of $\widehat{G}$;
Output: α^0, Z^0, β^0.

Intuitively speaking, the estimate of P by USVT is consistent when $\|P\|_*$ is "small". Following the arguments in Theorems 2.6 and 2.7 of [16], such a condition is satisfied when the covariate matrix $X = 0$ or when X has "simple" structure. Such "simple" structure could be $X_{ij} = \kappa(x_i, x_j)$ where $x_1, \cdots, x_n \in \mathbb{R}^d$ are feature vectors associated with the n nodes and $\kappa(\cdot, \cdot)$ characterizes the distance/similarity between node i and node j. For instance, one could have $X_{ij} = \mathbf{1}_{\{x_i = x_j\}}$ where $x_1, \cdots, x_n \in \{1, \cdots, K\}$ is a categorical variable such as gender, race, nationality, etc; or $X_{ij} = s(|x_i - x_j|)$ where $s(\cdot)$ is a continuous monotone link function and $x_1, \cdots, x_n \in \mathbb{R}$ is a continuous node covariate such as age, income, years of education, etc.

Remark 3.2. The computational cost of Algorithm 3 is dominated by matrix decompositions in step 1 (line 2) and step 3 (line 4). In the sparse case, the computation cost for the SVD part can be further reduced to be proportional to the number of edges in the sparse case.

4 Theoretical results

In this section, we first present error bounds for both fitting methods under inner-product models, followed by their generalizations to the more general models satisfying the Schoenberg condition (7). In addition, we give theoretical justifications of the two initialization methods for Algorithm 1.

4.1 Error bounds for inner-product models

We shall establish uniform high probability error bounds for inner-product models belonging to the following parameter space:

$$\mathcal{F}(n,k,M_1,M_2,X)=\Big\{\Theta|\Theta=\alpha 1_n{}^\top+1_n\alpha^\top+\beta X+ZZ^\top,\ JZ=Z,\\ \max_{1\leqslant i\leqslant n}\|Z_{i*}\|^2,\ \|\alpha\|_\infty,\ |\beta|\max_{1\leqslant i<j\leqslant n}|X_{ij}|\leqslant\frac{M_1}{3},\max_{1\leqslant i\neq j\leqslant n}\Theta_{ij}\leqslant -M_2\Big\}. \tag{16}$$

When $X=0$, we replace the first inequality in (16) with $\max_{1\leqslant i\leqslant n}\|Z_{i*}\|^2,\ \|\alpha\|_\infty\leqslant M_1/2$. For the results below, k, M_1, M_2 and X are all allowed to change with n.

Results for the convex approach We first present theoretical guarantees for the optimizer of (11). When X in nonzero, we make the following assumption for the identifiability of β.

Assumption 4.1. *The* stable rank *of the covariate matrix X satisfies* $\mathrm{r}_{\mathrm{stable}}(X)=\|X\|_{\mathrm{F}}^2/\|X\|_{\mathrm{op}}^2\geqslant M_0k$ *for some large enough constant M_0.*

The linear dependence on k of $\mathrm{r}_{\mathrm{stable}}(X)$ is in some sense necessary for β to be identifiable as otherwise the effect of the covariates could be absorbed into the latent component $ZZ^\top$.

Let $(\widehat{\alpha},\widehat{\beta},\widehat{G})$ be the solution to (11) and $(\alpha_\star,\beta_\star,G_\star)$ be the true parameter that governs the data generating process. Let $\widehat{\Theta}$ and $\Theta_\star$ be defined as in (2) but with the estimates and the true parameter values for the components respectively. Define the error terms $\Delta_{\widehat{\Theta}}=\widehat{\Theta}-\Theta_\star$, $\Delta_{\widehat{\alpha}}=\widehat{\alpha}-\alpha_\star$, $\Delta_{\widehat{\beta}}=\hat{\beta}-\beta_\star$ and $\Delta_{\widehat{G}}=\widehat{G}-G_\star$. The following theorem gives both deterministic and high probability error bounds for estimating both the latent vectors and logit-transformed probability matrix.

Theorem 4.1. *Under Assumption 4.1, for any λ_n satisfying $\lambda_n\geqslant\max\{2\|A-P\|_{\mathrm{op}},\ |\langle A-P,X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$, there exists a constant C such that*

$$\left\|\Delta_{\widehat{G}}\right\|_{\mathrm{F}}^2,\left\|\Delta_{\widehat{\Theta}}\right\|_{\mathrm{F}}^2\leqslant Ce^{2M_1}\lambda_n^2k.$$

Specifically, setting $\lambda_n=C_0\sqrt{\max\{ne^{-M_2},\log n\}}$ for a large enough positive constant C_0, there exist positive constants c,C such that uniformly over $\mathcal{F}(n,k,M_1,M_2,X)$, with probability at least $1-n^{-c}$,

$$\left\|\Delta_{\widehat{G}}\right\|_{\mathrm{F}}^2,\left\|\Delta_{\widehat{\Theta}}\right\|_{\mathrm{F}}^2\leqslant C\psi_n^2$$

where

$$\psi_n^2=e^{2M_1}nk\times\max\Big\{e^{-M_2},\frac{\log n}{n}\Big\}. \tag{17}$$

If we turn the error metrics in Theorem 4.1 to mean squared errors, namely $\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2/n^2$ and $\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2/n^2$, we obtain the familiar k/n rate in low rank matrix estimation problems and the theorem can viewed as a complementary result to the result in [22] in the case where there are observed covariate and the 1−bit matrix is fully observed. When $e^{-M_2} \geqslant \frac{\log n}{n}$, the sparsity of the network affects the rate through the multiplier e^{-M_2}. As the network gets sparser, the multiplier will be no smaller than $\frac{\log n}{n}$.

Remark 4.1. Note that the choice of the penalty parameter λ_n depends on e^{-M_2} where the maximum node degree of the network is of order $O(ne^{-M_2})$. In practice, we may not know this quantity and we propose to estimate M_2 with $\widehat{M}_2 = -\mathrm{logit}(\sum_{ij} A_{ij}/n^2)$.

Results for the non-convex approach A key step toward establishing the statistical properties of the outputs of Algorithm 1 is to characterize the evolution of its iterates. To start with, we introduce an error metric that is equivalent to $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 = \|\Theta^t - \Theta_\star\|_{\mathrm{F}}^2$ while at the same time is more convenient for establishing an inequality satisfied by all iterates. Note that the latent vectors are only identifiable up to an orthogonal transformation of $\mathbb{R}^k$, for any $Z_1, Z_2 \in \mathbb{R}^{n\times k}$, we define the distance measure

$$\mathrm{dist}(Z_1, Z_2) = \min_{R\in O(k)} \|Z_1 - Z_2 R\|_{\mathrm{F}}$$

where $O(k)$ collects all $k \times k$ orthogonal matrices. Let $R^t = \arg\min_{R\in O(k)} \left\|Z^t - Z_\star R\right\|_{\mathrm{F}}$ and $\Delta_{Z^t} = Z^t - Z_\star R^t$, and further let $\Delta_{\alpha^t} = \alpha^t - \alpha_\star, \Delta_{G^t} = Z^t(Z^t)^\top - Z_\star Z_\star^\top$ and $\Delta_{\beta^t} = \beta^t - \beta_\star$. Then the error metric we use for characterizing the evolution of iterates is

$$e_t = \|Z_\star\|_{\mathrm{op}}^2 \|\Delta_{Z^t}\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2. \tag{18}$$

Let $\kappa_{Z_\star}$ be the condition number of $Z_\star$ (i.e., the ratio of the largest to the smallest singular values). The following lemma shows that the two error metrics e_t and $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2$ are equivalent up to a multiplicative factor of order $\kappa_{Z_\star}^2$.

Lemma 4.1. *Under Assumption 4.1, there exists a constant* $0 \leqslant c_0 < 1$ *such that*

$$e_t \leqslant \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)} \|\Delta_{G^t}\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2 \leqslant \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)(1-c_0)} \|\Delta_{\Theta^t}\|_{\mathrm{F}}^2.$$

Moreover, if dist$(Z^t, Z_\star) \leqslant c\|Z_\star\|_{\mathrm{op}}$,

$$e_t \geqslant \frac{1}{(c+2)^2} \|\Delta_{G^t}\|_{\mathrm{F}}^2 + 2\left\|\Delta_{\alpha^t} 1_n{}^\top\right\|_{\mathrm{F}}^2 + \left\|\Delta_{\beta^t} X\right\|_{\mathrm{F}}^2 \geqslant \frac{1}{(c+2)^2(1+c_0)} \|\Delta_{\Theta^t}\|_{\mathrm{F}}^2.$$

In addition, our error bounds depend on the following assumption on the initializers.

Assumption 4.2. *The initializers* Z^0, α^0, β^0 *in Algorithm 1 satisfy* $e_0 \leqslant ce^{-2M_1} \|Z_\star\|_{\mathrm{op}}^4 / \kappa_{Z_\star}^4$ *for a sufficiently small positive constant c.*

Note that the foregoing assumption is not very restrictive. If $k \ll n$, M_1 is a constant and the entries of $Z_\star$ are i.i.d. random variables with mean zero and bounded variance, then $\|Z_\star\|_{\mathrm{op}} \asymp \sqrt{n}$ and $\kappa_{Z_\star} \asymp 1$. In view of Lemma 4.1, this only requires $\frac{1}{n^2}\|\Theta^0 - \Theta_\star\|_{\mathrm{F}}^2$ to be upper bounded by some

constant. We defer verification of this assumption for initial estimates constructed by Algorithm 2 and Algorithm 3 to Section 4.3.

The following theorem states that under such an initialization, errors of the iterates converge linearly till they reach the same statistical precision ψ_n^2 as in Theorem 4.1 modulo a multiplicative factor that depends only on the condition number of $Z_\star$.

Theorem 4.2. *Let Assumptions 4.1 and 4.2 be satisfied. Set the constraint sets as*[1]

$$\mathcal{C}_Z = \{Z \in \mathbb{R}^{n\times k}, JZ = Z, \max_{1\leqslant i\leqslant n} \|Z_{i*}\| \leqslant M_1/3\},$$
$$\mathcal{C}_\alpha = \{\alpha \in \mathbb{R}^n, \|\alpha\|_\infty \leqslant M_1/3\},\ \mathcal{C}_\beta = \{\beta \in \mathbb{R}, \beta\|X\|_\infty \leqslant M_1/3\}.$$

Set the step sizes as in (13) *for any* $\eta \leqslant c$ *where* $c > 0$ *is a universal constant. Let* $\zeta_n = \max\{2\|A - P\|_{\mathrm{op}},\ |\langle A - P, X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$. *Then we have*

1. Deterministic errors of iterates: *If* $\|Z_\star\|_{\mathrm{op}}^2 \geqslant C_1\kappa_{Z_\star}^2 e^{M_1}\zeta_n^2 \times \max\left\{\sqrt{\eta k e^{M_1}}, 1\right\}$ *for a sufficiently large constant* C_1, *there exist positive constants* ρ *and* C *such that*

$$e_t \leqslant 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{C\kappa_{Z_\star}^2}{\rho} e^{2M_1}\zeta_n^2 k.$$

2. Probabilistic errors of iterates: *If* $\|Z_\star\|_{\mathrm{op}}^2 \geqslant C_1\kappa_{Z_\star}^2\sqrt{n}e^{M_1 - M_2/2}\max\left\{\sqrt{\eta k e^{M_1}}, 1\right\}$ *for a sufficiently large constant* C_1, *there exist positive constants* ρ, c_0 *and* C *such that uniformly over* $\mathcal{F}(n, k, M_1, M_2, X)$ *with probability at least* $1 - n^{-c_0}$,

$$e_t \leqslant 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + C\frac{\kappa_{Z_\star}^2}{\rho}\psi_n^2.$$

For any $T > T_0 = \log(\frac{M_1^2}{\kappa_{Z_\star}^2 e^{4M_1 - M_2}}\frac{n}{k^2})/\log(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho)$,

$$\|\Delta_{G^T}\|_{\mathrm{F}}^2,\ \|\Delta_{\Theta^T}\|_{\mathrm{F}}^2 \leqslant C'\kappa_{Z_\star}^2\psi_n^2.$$

for some constant $C' > 0$.

Remark 4.2. In view of Lemma 4.1, the rate obtained by the non-convex approach in terms of $\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2$ matches the upper bound achieved by the convex method, up to a multiplier of squared condition number $\kappa_{Z_\star}^2$. As suggested by Lemma 4.1, the extra factor comes partly from the fact that e_t is a slightly stronger loss function than $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2$ and in the worst case can be $c\kappa_{Z_\star}^2$ times larger than $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2$.

Remark 4.3. Under the setting of Theorem 4.2, the projection steps for α, β in Algorithm 1 are straightforward and have the following closed form expressions: $\alpha_i^{t+1} = \widetilde{\alpha}_i^{t+1}\min\{1, M_1/(3|\widetilde{\alpha}_i^{t+1}|)\}$,

[1]When $X = 0$, $\mathcal{C}_\beta = \varnothing$ and we replace $M_1/3$ in $\mathcal{C}_Z$ and $\mathcal{C}_\alpha$ with $M_1/2$ in correspondence with the discussion following (16).

$\beta^{t+1} = \widetilde{\beta}^{t+1} \min\{1, M_1/(3|\widetilde{\beta}^{t+1}| \max_{i,j} |X_{ij}|)\}$. The projection step for Z is slightly more involved. Notice that $\mathcal{C}_Z = \mathcal{C}_Z^1 \bigcap \mathcal{C}_Z^2$ where

$$\mathcal{C}_Z^1 = \{Z \in \mathbb{R}^{n\times k}, JZ = Z\},\ \mathcal{C}_Z^2 = \{Z \in \mathbb{R}^{n\times k}, \max_{1\leqslant i\leqslant n} \|Z_{i*}\|^2 \leqslant M_1/3\}.$$

Projecting to either of them has closed form solution, that is $\mathcal{P}_{\mathcal{C}_Z^1}(Z) = JZ$, $[\mathcal{P}_{\mathcal{C}_Z^2}(Z)]_{i*} = Z_{i*} \min\{1, \sqrt{M_1/(3\|Z_{i*}\|^2)}\}$. Then Dykstra's projection algorithm [23] (or alternating projection algorithm) can be applied to obtain $\mathcal{P}_{\mathcal{C}_Z}(\widetilde{Z}^{t+1})$. We note that projections induced by the boundedness constraints for Z, α, β are needed for establishing the error bounds theoretically. However, when implementing the algorithm, users are at liberty to drop these projections and to only center the columns of the Z iterates as in (14). We did not see any noticeable difference thus caused on simulated examples reported in Section 5.

Remark 4.4. When both M_1 and M_2 are constants and the covariate matrix X is absent, the result in Section 4.5 of [18], in particular Corollary 5, implies the error rate of $O(nk)$ in Theorem 4.2. However, when $M_1 \to \infty$ and M_2 remains bounded as $n \to \infty$, the error rate in [18] becomes[2] $O(e^{8M_1} M_1^2 nk)$, which can be much larger than the rate $O(e^{2M_1} nk)$ in Theorem 4.2 even when X is absent. We feel that this is a byproduct of the pursuit of generality in [18] and so the analysis has not been fine-tuned for latent space models. In addition, Algorithm 1 enjoys nice theoretical guarantees on its performance even when the model is mis-specified and the Θ matrix is only approximately low rank. See Theorem 4.4 below. This case which is important from a modeling viewpoint was not considered in [18] as its focus was on generic non-convex estimation of low rank positive semi-definite matrices rather than fittings latent space models.

4.2 Error bounds for more general models

We now investigate the performances of the fitting approaches on more general models satisfying the Schoenberg condition (7). To this end, we consider the following parameter space for the more general class of latent space models

$$\begin{aligned}\mathcal{F}_g(n, M_1, M_2, X) = \Big\{\Theta | \Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \beta X + G, G \in \mathbb{S}_+^n, JG = G, \\ \max_{1\leqslant i\leqslant n} G_{ii},\ \|\alpha\|_\infty,\ |\beta| \max_{1\leqslant i<j\leqslant n} |X_{ij}| \leqslant M_1/3, \max_{1\leqslant i\neq j\leqslant n} \Theta_{ij} \leqslant -M_2\Big\}.\end{aligned} \tag{19}$$

As before, when $X = 0$, we replace the first inequality in (19) with $\max_{1\leqslant i\leqslant n} \|Z_{i*}\|^2$, $\|\alpha\|_\infty \leqslant M_1/2$. For the results below, M_1, M_2 and X are all allowed to change with n. Note that the latent space dimension k is no longer a parameter in (19). Then for any positive integer k, let $U_k D_k U_k^\top$ be the best rank-k approximation to $G_\star$. In this case, with slight abuse of notation, we let

$$Z_\star = U_k D_k^{1/2} \quad \text{and} \quad \bar{G}_k = G_\star - U_k D_k U_k^\top. \tag{20}$$

Note that (19) does not specify the spectral behavior of G which will affect the performance of the fitting methods as the theorems in this section will later reveal. We choose not to make such specification due to two reasons. First, the spectral behavior of distance matrices resulting from

[2] One can verify that in this case we can identify the quantities in Corollary 5 of [18] as $\sigma = 1$, $p = 1$, $d = n$, $r = k$, $\nu \asymp M_1$, $L_{4\nu} \asymp 1$ and $\ell_{4\nu} \asymp e^{4M_1}$.

different kernel functions and manifolds is by itself a very rich research topic. See, for instance, [48, 8, 24, 20] and the references therein. In addition, the high probability error bounds we are to develop in this section is going to work uniformly for all models in (19) and can be specialized to any specific spectral decay pattern of G of interest.

Results for the convex approach The following theorem is a generalization of Theorem 4.1 to the general class (19).

Theorem 4.3. *For any $k \in \mathbb{N}_+$ such that Assumption 4.1 holds and any λ_n satisfying $\lambda_n \geqslant \max\{2\|A-P\|_{\mathrm{op}},\ |\langle A-P, X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$, there exists a constant C such that the solution to the convex program (11) satisfies*

$$\left\|\Delta_{\widehat{\Theta}}\right\|_{\mathrm{F}}^2 \leqslant C\left(e^{2M_1}\lambda_n^2 k + e^{M_1}\lambda_n\|\bar{G}_k\|_*\right).$$

Specifically, setting $\lambda_n = C_0\sqrt{\max\{ne^{-M_2}, \log n\}}$ for a large enough constant C_0, there exists positive constants c, C such that uniformly over $\mathcal{F}_g(n, M_1, M_2, X)$ with probability at least $1-n^{-c}$,

$$\left\|\Delta_{\widehat{\Theta}}\right\|_{\mathrm{F}}^2 \leqslant C(\psi_n^2 + e^{M_1-M_2/2}\sqrt{n}\|\bar{G}_k\|_*). \tag{21}$$

The upper bound in (21) has two terms. The first is the same as that for the inner-product model. The second can be understood as the effect of model mis-specification, since the estimator is essentially based on the log-likelihood of the inner-product model. We note that the bound holds for any k such that Assumption 4.1 holds while the choice of the tuning parameter λ_n does not depend on k. Therefore, we can take the infimum over all admissible values of k, depending on the stable rank of X. When $X = 0$, we can further improve the bound on the right side of (21) to be the infimum of the current expression over all $0 \leqslant k \leqslant n$.

Results for the non-convex approach Under the definition in (20), we continue to use the error metric e_t defined in equation (18). The following theorem is a generalization of Theorem 4.2 to the general class (19).

Theorem 4.4. *Under Assumptions 4.1 and 4.2, set the constraint sets $\mathcal{C}_Z$, $\mathcal{C}_\alpha$, $\mathcal{C}_\beta$ and the step sizes η_Z, η_α and η_β as in Theorem 4.2. Let $\zeta_n = \max\{2\|A-P\|_{\mathrm{op}},\ |\langle A-P, X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$. Then we have*

1. Deterministic errors of iterates: *If $\|G_\star\|_{\mathrm{op}} \geqslant C_1\kappa_{Z_\star}^2 e^{M_1}\zeta_n^2 \times \max\left\{\sqrt{\eta k e^{M_1}}, \sqrt{\eta\|\bar{G}_k\|_{\mathrm{F}}^2/\zeta_n^2}, 1\right\}$, there exist positive constants ρ and C such that*

$$e_t \leqslant 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{C\kappa_{Z_\star}^2}{\rho}\left(e^{2M_1}\zeta_n^2 k + e^{M_1}\left\|\bar{G}_k\right\|_{\mathrm{F}}^2\right).$$

2. Probabilistic errors of iterates: *If $\|G_\star\|_{\mathrm{op}} \geqslant C_1\kappa_{Z_\star}^2\sqrt{n}e^{M_1-M_2/2}\max\left\{\sqrt{\eta k e^{M_1}}, \sqrt{\eta\|\bar{G}_k\|_{\mathrm{F}}^2/\zeta_n^2}, 1\right\}$ for a sufficiently large constant C_1, there exist positive constants ρ, c_0 and C such that*

uniformly over $\mathcal{F}_g(n, M_1, M_2, X)$ *with probability at least* $1 - n^{-c_0}$*, the iterates generated by Algorithm 1 satisfying*

$$e_t \leqslant 2\left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{C\kappa_{Z_\star}^2}{\rho}\left(\psi_n^2 + e^{M_1}\left\|\bar{G}_k\right\|_{\mathrm{F}}^2\right).$$

For any $T > T_0 = \log(\frac{M_1^2}{\kappa_{Z_\star}^2 e^{4M_1 - M_2}}\frac{n}{k^2})/\log(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho)$,

$$\left\|\Delta_{G^T}\right\|_{\mathrm{F}}^2,\ \left\|\Delta_{\Theta^T}\right\|_{\mathrm{F}}^2 \leqslant C'\kappa_{Z_\star}^2\left(\psi_n^2 + e^{M_1}\left\|\bar{G}_k\right\|_{\mathrm{F}}^2\right)$$

for some constant C'*.*

Compared with Theorem 4.2, the upper bound here has an additional term $e^{M_1}\|\bar{G}_k\|_{\mathrm{F}}^2$ that can be understood as the effect of model mis-specification. Such a term can result from mis-specifying the latent space dimension in Algorithm 1 when the inner-product model holds, or it can arise when the inner-product model is not the true underlying data generating process. This term is different from its counterpart in Theorem 4.3 which depends on $\bar{G}_k$ through its nuclear norm. In either case, the foregoing theorem guarantees that Algorithm 1 continues to yield reasonable estimate of Θ and G as long as $\|\bar{G}_k\|_{\mathrm{F}}^2 = O(e^{M_1 - M_2}nk)$, i.e., when the true underlying model can be reasonably approximated by an inner-product model with latent space dimension k.

4.3 Error bounds for initialization

We conclude this section with error bounds for the two initialization methods in Section 3. These bounds justify that the methods yield initial estimates satisfying Assumption 4.2 under different circumstances.

Error bounds for Algorithm 2 The following theorem indicates that Algorithm 2 yields good initial estimates after a small number of iterates as long as the latent effect $G_\star$ is substantial and the remainder $\bar{G}_k$ is well controlled.

Theorem 4.5. *Suppose that Assumption 4.1 holds and that* $\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \|\beta_\star X\|_{\mathrm{F}} \leqslant C\|G_\star\|_{\mathrm{F}}$ *for a numeric constant* $C > 0$*. Let* λ_n *satisfy* $C_0\sqrt{\max\{ne^{-M_2}, \log n\}} \leqslant \lambda_n \leqslant c_0\|G_\star\|_{\mathrm{op}}/(e^{2M_1}\sqrt{k}\kappa_{Z_\star}^3)$ *for sufficiently large constant* C_0 *and sufficiently small constant* c_0*, let* γ_n *satisfy* $\gamma_n \leqslant \delta\lambda_n/\|G_\star\|_{\mathrm{op}}$ *for sufficiently small constant* δ*. Choose step size* $\eta \leqslant 2/9$ *and set the constraint sets as* [3]

$$\mathcal{C}_G = \{G \in \mathbb{S}_+^{n\times n}, JG = G, \max_{1\leqslant i,j\leqslant n}|G_{ij}| \leqslant M_1/3\},$$
$$\mathcal{C}_\alpha = \{\alpha \in \mathbb{R}^n, \|\alpha\|_\infty \leqslant M_1/3\},\ \mathcal{C}_\beta = \{\beta \in \mathbb{R}, \beta\|X\|_\infty \leqslant M_1/3\}.$$

If the latent vectors contain strong enough signal in the sense that

$$\|G_\star\|_{\mathrm{op}}^2 \geqslant C\kappa_{Z_\star}^6 e^{2M_1}\max\left\{\phi_n^2,\ \|\bar{G}_k\|_*^2/k,\ \left\|\bar{G}_k\right\|_{\mathrm{F}}^2\right\}, \tag{22}$$

[3]When $X = 0$, $\mathcal{C}_\beta = \varnothing$ and we replace $M_1/3$ in $\mathcal{C}_G$ and $\mathcal{C}_\alpha$ with $M_1/2$ in correspondence with the discussion following (16) and (19).

for some sufficiently large constant C, there exist positive constants c, C_1 such that with probability at least $1 - n^{-c}$, for any given constant $c_1 > 0$, $e_T \leqslant c_1^2 e^{-2M_1} \|Z_\star\|_{\mathrm{op}}^4 / \kappa_{Z_\star}^4$ as long as $T \geqslant T_0$, where

$$T_0 = \log\left(\frac{C_1 e^{2M_1} k \kappa_{Z_\star}^6}{c_1^2}\right)\left(\log\left(\frac{1}{1-\gamma_n\eta}\right)\right)^{-1}. \tag{23}$$

We note that the theorem holds for both inner-product models and more general models satisfying condition (7). In addition, it gives the ranges of λ_n and γ_n for implementing Algorithm 2. Note that the choices λ_n and γ_n affect the number of iterations needed. To go one step further, the following corollary characterizes the ideal choices of γ_n and λ_n in Algorithm 2. It is worth noting that the choice of λ_n here does not coincide with that in Theorem 4.1 and Theorem 4.3. So this is slightly different from the conventional wisdom that to initialize the non-convex approach, it would suffice to simply run the convex optimization algorithm for a small number of steps. Interestingly, the corollary shows that when M_1, k and $\kappa_{Z_\star}$ are all upper bounded by universal constants, for appropriate choices of γ_n and λ_n in Algorithm 2, the number of iterations needed does not depend on the graph size n.

Corollary 4.1. *Specifically in Theorem 4.5, if we choose $\gamma_n = c_0/(e^{2M_1}\sqrt{k}\kappa_{Z_\star}^3)$ for some sufficiently small constant c_0, and $\lambda_n = C_0\gamma_n \|G_\star\|_{\mathrm{op}}$ for some sufficiently large constant C_0, there exist positive constants c, C_1 such that with probability at least $1 - n^{-c}$, for any given constant $c_1 > 0$, $e_T \leqslant c_1^2 e^{-2M_1} \|Z_\star\|_{\mathrm{op}}^4 / \kappa_{Z_\star}^4$ as long as $T \geqslant T_0$, where*

$$T_0 = \log\left(\frac{C_1 e^{2M_1} k \kappa_{Z_\star}^6}{c_1^2}\right)\left(\log\left(\frac{1}{1-\gamma_n\eta}\right)\right)^{-1}. \tag{24}$$

Remark 4.5. Similar to computing $\mathcal{P}_{\mathcal{C}_Z}(\cdot)$ in Algorithm 1, $\mathcal{P}_{\mathcal{C}_G}(\cdot)$ could also be implemented by Dykstra's projection algorithm since $\mathcal{C}_G$ is the intersection of two convex sets. The boundedness constraint $\max_{i,j}|G_{ij}| \leqslant M/3$ is only for the purpose of proof. In practice, if ignoring this constraint, G^{t+1} will have closed form solution $G^{t+1} = \mathcal{P}_{\mathbb{S}_+^n}(J\widetilde{G}^{t+1}J)$ where $\mathcal{P}_{\mathbb{S}_+^n}(\cdot)$ can be computed by singular value thresholding.

Error bounds for Algorithm 3 The following result justifies the singular value thresholding approach to initialization for inner-product models with no edge covariate.

Proposition 4.1. *If no covariates are included in the latent space model and $\|G_\star\|_{\mathrm{F}} \geqslant c_0 n$ for some numeric constant $c_0 > 0$, then there exists constant c_1 such that with probability at least $1 - n^{c_1}$, for any $n \geqslant C(k, M_1, \kappa_{Z_\star})$ where $C(k, M_1, \kappa_{Z_\star})$ is a constant depending on k, M_1 and $\kappa_{Z_\star}$, the outputs of Algorithm 3 with $\tau \geqslant 1.1\sqrt{n}$ satisfies the initialization condition in Assumption 4.2.*

Although we do not have further theoretical results, Algorithm 3 worked well on all the simulated data examples reported in Section 5.

5 Simulation studies

In this section, we present results of simulation studies on three different aspects of the proposed methods: (1) scaling of estimation errors with network sizes, (2) impact of initialization on Algorithm 1, and (3) performance of the methods on general models.

Estimation errors We first investigate how estimation errors scale with network size. To this end, we fix $\beta_\star = -\sqrt{2}$ and for any $(n, k) \in \{500, 1000, 2000, 4000, 8000\} \times \{2, 4, 8\}$, we set the other model parameters randomly following these steps:

1. Generate the degree heterogeneity parameters: $(\alpha_\star)_i = -\alpha_i / \sum_{j=1}^n \alpha_j$ for $1 \leqslant i \leqslant n$, where $\alpha_1, \cdots, \alpha_n \overset{iid}{\sim} U[1, 3]$.
2. Generate $\mu_1,\ \mu_2 \in \mathbb{R}^k$ with coordinates iid following $U[-1, 1]$ as two latent vector centers;
3. Generate latent vectors: for $i = 1, \ldots, k$, let $(z_1)_i, \cdots, (z_{\lfloor n/2 \rfloor})_i \overset{iid}{\sim} (\mu_1)_i + N_{[-2,2]}(0, 1)$ and $(z_{\lfloor n/2 \rfloor + 1})_i, \cdots, (z_n)_i \overset{iid}{\sim} (\mu_2)_i + N_{[-2,2]}(0, 1)$ where $N_{[-2,2]}(0, 1)$ is the standard normal distribution restricted onto the interval $[-2, 2]$, then set $Z_\star = JZ$ where $Z = [z_1, \cdots, z_n]^\top$ and J is as defined in (3). Finally, we normalize $Z_\star$ such that $\|G_\star\|_{\mathrm{F}} = n$;
4. Generate the covariate matrix: $X = n\widetilde{X}/\|\widetilde{X}\|_{\mathrm{F}}$ where $\widetilde{X}_{ij} \overset{iid}{\sim} \min\{|N(1, 1)|, 2\}$.

For each generated model, we further generated 30 independent copies of the adjacency matrix for each model configuration. Unless otherwise specified, for all experiments in this section, with given (n, k), the model parameters are set randomly following the above four steps and algorithms are run on 30 independent copies of the adjacency matrix.

The results of the estimation error for varying (n, k) are summarized in the log-log boxplots in Figure 1, where "Relative Error - Z" is defined as $\|\widehat{Z}\widehat{Z}^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 / \|Z_\star Z_\star^\top\|_{\mathrm{F}}^2$ and "Relative Error - Θ" is defined as $\|\widehat{\Theta} - \Theta_\star\|_{\mathrm{F}}^2 / \|\Theta_\star\|_{\mathrm{F}}^2$. From the boxplots, for each fixed latent space dimension k, the relative estimation errors for both $Z_\star$ and $\Theta_\star$ scale at the order of $1/\sqrt{n}$. This agrees well with the theoretical results in Section 3. For different latent space dimension k, the error curve (in log-log scale) with respect to network size n only differs in the intercept.

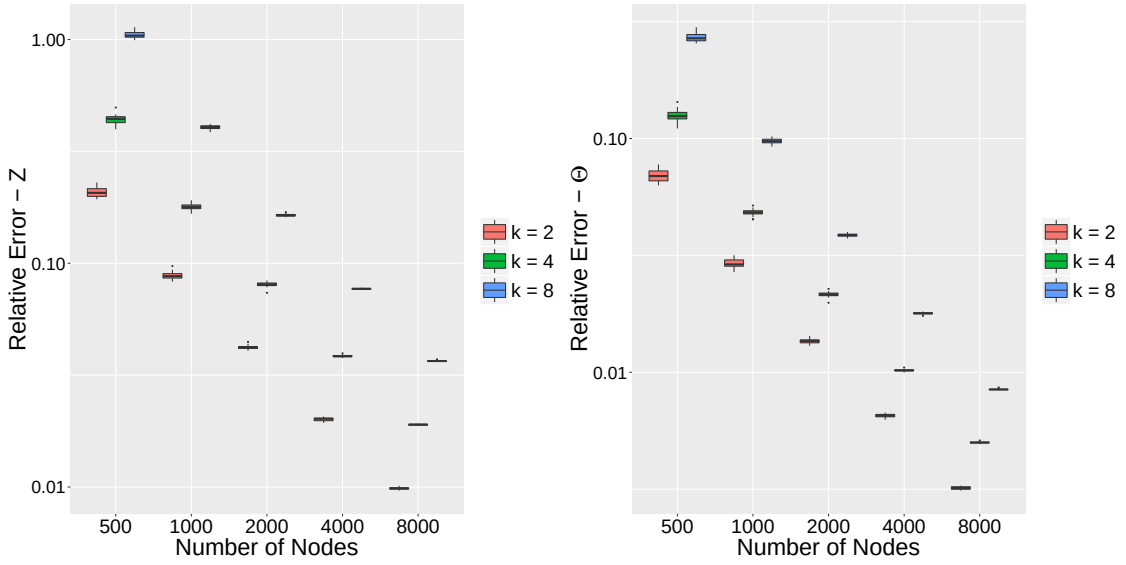

Figure 1: log-log boxplot for relative estimation errors with varying network size and latent space dimension.

Impact of initialization on Algorithm 1 We now turn to the comparison of three different initialization methods for Algorithm 1: the convex method (Algorithm 2), singular value thresholding (Algorithm 3), and random initialization. To this end, we fixed $n = 4000, k = 4$. In Algorithm 2, we choose $T = 10$ and $\lambda_n = 2\sqrt{n\widehat{p}}$ where $\widehat{p} = \sum_{ij} A_{ij}/n^2$. In Algorithm 3, we set $M_1 = 4$ and threshold $\tau = \sqrt{n\widehat{p}}$. The relative estimation errors are summarized as boxplots in Figure 2. Clearly, the non-convex algorithm is very robust to the initial estimates. Similar phenomenon is observed in real data analysis where different initializations yield nearly the same clustering accuracy.

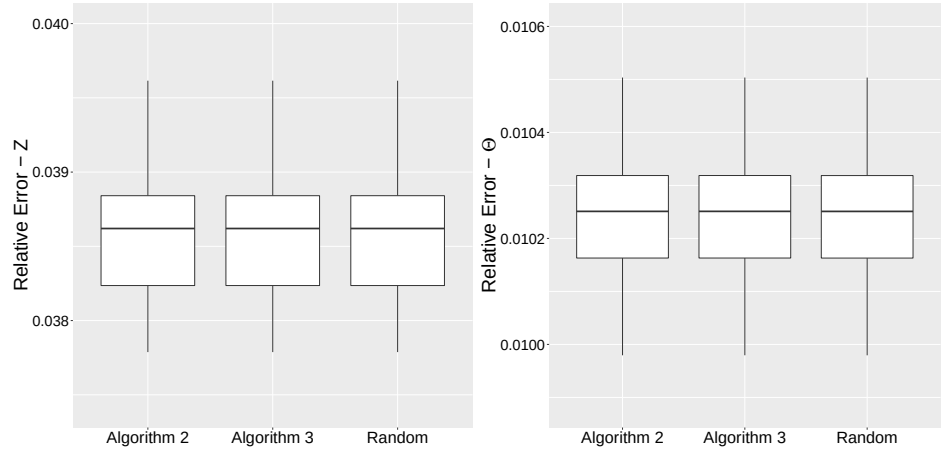

Figure 2: Boxplot for relative estimation error with different initialization methods.

Performance on general models To investigate the performance of the proposed method under the general model (4), we try two frequently used kernel functions, distance kernel $\ell_d(z_i, z_j) = -\|z_i - z_j\|$ and Gaussian kernel $\ell_g(z_i, z_j) = 4\exp(-\|z_i - z_j\|^2/9)$. In this part, we use d to represent the dimension of the latent vectors (that is, $z_1, \cdots, z_n \in \mathbb{R}^d$) and k to represent the fitting dimension in Algorithm 1. We fix $d = 4$ and network size $n = 4000$. Model parameters are set randomly in the same manner as the four step procedure except that the third step is changed to:

> Generate latent vectors: for $i = 1, \ldots, d$, let $(z_1)_i, \cdots, (z_{\lfloor n/2 \rfloor})_i \overset{iid}{\sim} (\mu_1)_i + N_{[-2,2]}(0, 1)$ and $(z_{\lfloor n/2 \rfloor + 1})_i, \cdots, (z_n)_i \overset{iid}{\sim} (\mu_2)_i + N_{[-2,2]}(0, 1)$ where $N_{[-2,2]}(0, 1)$ is the standard normal distribution restricted onto the interval $[-2, 2]$. Finally for given kernel function $\ell(\cdot, \cdot)$, set $G_\star = JLJ$ where $L_{ij} = \ell(z_i, z_j)$.

We run both the convex approach and Algorithm 1 with different fitting dimensions. The boxplot for the relative estimation errors and the singular value decay of the kernel matrix under distance kernel and Gaussian kernel are summarized in Figure 3 and Figure 4 respectively.

As we can see, under the generalized model, the non-convex algorithm exhibits bias-variance tradeoff with respect to the fitting dimension, which depends on the singular value decay of the kernel matrix. The advantage of the convex method is the adaptivity to the unknown kernel function.

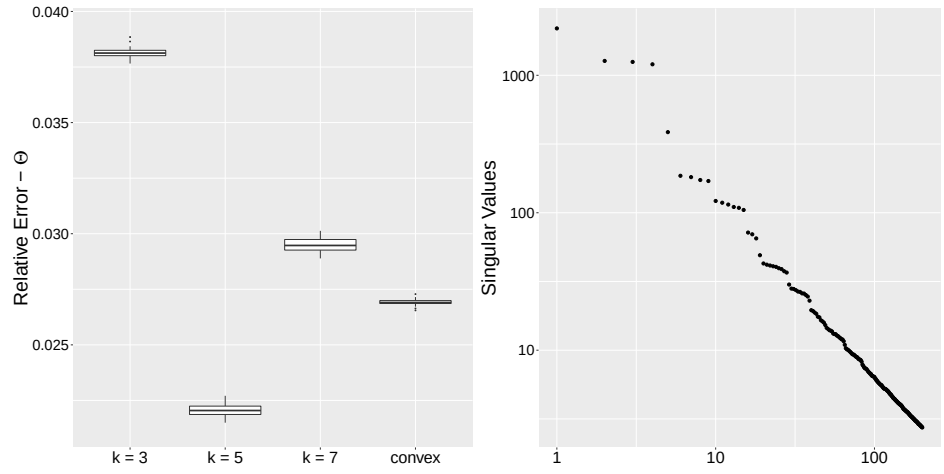

Figure 3: The log-log plot of relative estimation errors of both convex and non-convex approach under the distance kernel $\ell_d(z_i, z_j) = -\|z_i - z_j\|$ (left panel). The log-log plot of ordered eigenvalues of $G_\star$ (right panel).

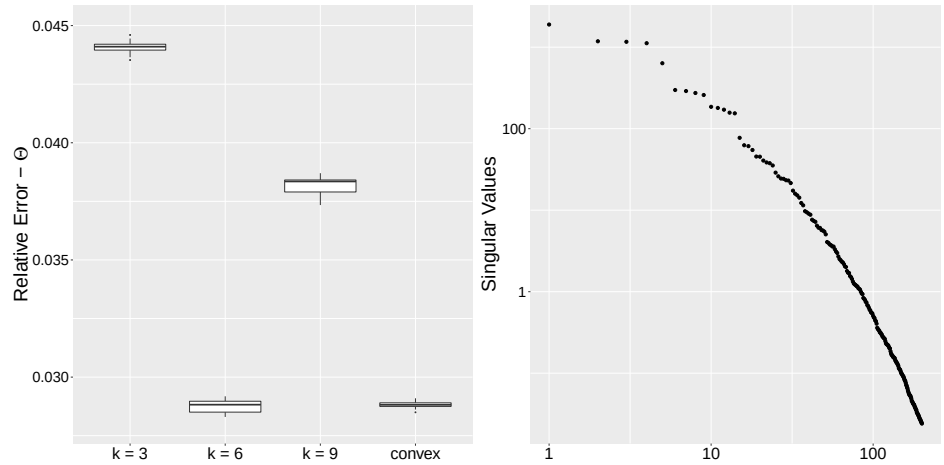

Figure 4: log-log plot for the relative estimation errors of both convex and non-convex approach under the Gaussian kernel $\ell_g(z_i, z_j) = 4\exp(-\|z_i - z_j\|^2/9)$ (left panel). The log-log plot of ordered eigenvalues of $G_\star$ (right panel).

When the true underlying model is not the inner-product model, Theorem 4.4 indicates that the optimal choice of fitting dimension k should depend on the size of the network. To illustrate such dependency, we vary both network size and fitting dimension, of which the results are summarized in Figure 5. As the size of the network increases, the optimal choice of fitting dimension increases as well.

Computational cost and scalability Finally, to test the scalability of both non-convex and convex algorithms, we also record the runtimes of the simulation for different sizes of the network and different dimensions of the latent vectors. The left and right panels of Figure 6 summarize the

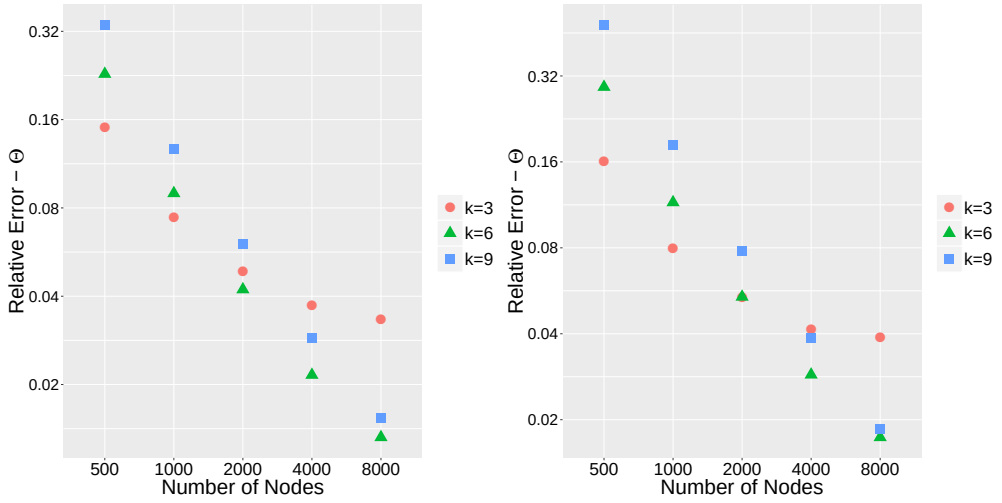

Figure 5: log-log plot for the relative estimation error with varying network size under distance kernel $\ell_d(z_i, z_j) = -\|z_i - z_j\|$ (left panel) and under the Gaussian kernel $\ell_g(z_i, z_j) = 4\exp(-\|z_i - z_j\|^2/9)$ (right panel).

runtimes for the non-convex and convex algorithms respectively. For the convex algorithm we set $\lambda_n = 2\sqrt{n\widehat{p}}$ with $\widehat{p} = \sum_{ij} A_{ij}/n^2$, and for the non-convex method we use true values of the latent space dimension. As the plots suggest, the relationship between runtimes and number of nodes is approximately linear on a log-log scale. The slopes of the two plots reveal that the runtime is close to $O(n^2)$ up to some poly-logarithmic multiplier for the non-convex algorithm, and close to $O(n^3)$ for the convex algorithm. Furthermore, the runtimes do not seem sensitive to the latent space dimension.

An obvious algorithmic competitor to consider is the Bayes fitting method proposed in [31, 32]. Due to the Bayes nature of the method, it runs quite a bit slower than the present two algorithms. However, such a comparison is unfair as the Bayes method provides substantial additional information of the posterior distribution, such as credible sets.

6 Real data examples

In this section, we demonstrate how the model and fitting methods can be used to explore real world datasets that involve large networks. In view of the discussion in Section 4.2 and Section 5, we can always employ the inner-product model (1) as the working model. In particular, we illustrate three different aspects. First, we consider community detection on networks without covariate. To this end, we compare the performance of simple k-means clustering on fitted latent variables with several state-of-the-art methods. Next, we investigate community detection on networks with covariates. In this case, we could still apply k-means clustering on fitted latent variables. Whether there is covariate or not, we can always visualize the network by plotting fitted latent variables in some appropriate way. Furthermore, we study how fitting the model can generate new feature variables to aid content-based classification of documents. The ability of feature generation also makes the model and the fitting methods potentially useful in other learning scenarios when additional

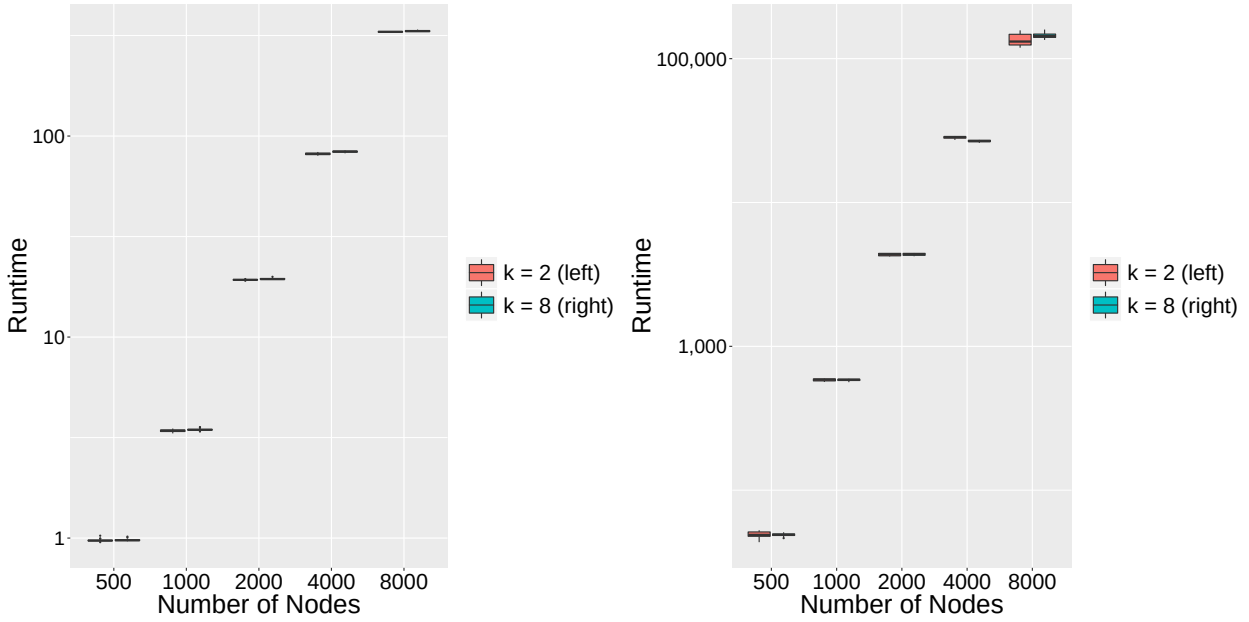

Figure 6: The log-log plot of runtimes (in seconds) for the non-convex and convex approaches.

network information among both training and test data is present.

6.1 Community detection without covariate

Community detection on networks without covariate has been intensively studied from both theoretical and methodological viewpoints. Thus, it naturally serves as a test example for the effectiveness of the model and fitting methods we have proposed in previous sections. To adapt our method to community detection, we propose to partition the nodes by the following two step procedure:

1. Fit the inner-product model to data with Algorithm 1;
2. Apply a simple k-means clustering on the fitted latent variables.

In what follows, we call this two step procedure LSCD (Latent Space based Community Detection), and in all the examples below, we used Algorithm 3 for initialization of Algorithm 1. We shall compare it with four state-of-the-art methods: (1) SCORE [35]: a normalized spectral clustering method developed under degree-corrected block models (DCBM); (2) OCCAM [64]: a normalized and regularized spectral clustering method for potentially overlapping community detection; (3) CMM [19]: a convexified modularity maximization method developed under DCBM. (4) Latentnet [42]: a hierachical bayesian method based on the latent space clustering model [30]. For theoretical work on latent space model based community detection, see [52] which provided some theory when the minimum node degree grows at a rate $\frac{n}{\sqrt{\log n}}$.

To avoid biasing toward our own method, we compare these methods on three datasets that have been previously used in the original papers to justify the first three methods at comparison: a political blog dataset [1] that was studied in [35] and two Facebook datasets (friendship networks

of Simmons College and Caltech) [59] that were studied in [19]. To make fair comparison, for all the methods, we supplied the true number of communities in each data. When fitting our model, we set the latent space dimension to be the same as the number of communities.

In the latentnet package [42], there are three different ways to predict the community membership. Using the notation of the R package [42], they are `mkl$Z.K`, `mkl$mbc$Z.K` and `mle$Z.K`. We found that `mkl$mbc$Z.K` consistently outperformed the other two on these data examples and we thus used it as the outcome of Latentnet. Due to the stochastic nature of the Bayesian approach, we repeated it 20 times on each dataset and reported both the average errors as well as the standard deviations (numbers in parentheses).

Table 1 summarizes the performance of all five methods on the three datasets. For columns LSCD, SCORE and OCCAM, we set the latent space dimension or the number of eigenvectors of the corresponding methods at k, where k is the number of clusters to seek in the data. Following a referee's suggestion, for columns LSCD_{k+1}, SCORE_{k+1} and OCCAM_{k+1}, we set the latent space dimension or the number of eigenvectors at $k+1$. Among all the methods at comparison, all methods except SCORE_{k+1} and OCCAM_{k+1} performed well on the political blog dataset with Latentnet being the best, and LSCD outperformed all other methods on the two Facebook datasets. Moreover, while SCORE and OCCAM get improvement in performance by using one more eigenvector, LSCD is less sensitive to the choice of tuning parameter.

Dataset	# Clusters	LSCD	SCORE	OCCAM	CMM
Political Blog	2	4.746%	4.746%	5.319%	5.074%
Simmons College	4	11.79%	23.57%	22.43%	12.04%
Caltech	8	**17.97%**	30.34%	31.19%	21.02%
Dataset	# Clusters	LSCD_{k+1}	SCORE_{k+1}	OCCAM_{k+1}	Latentnet
Political Blog	2	4.583%	23.159%	7.201%	**4.513**% (0.117%)
Simmons College	4	**10.99%**	16.45%	15.13%	29.09% (1.226%)
Caltech	8	18.98%	25.42%	23.22%	38.47% (1.190%)

Table 1: A summary on proportions of mis-clustered nodes by different methods on three datasets.

In what follows, we provide more details on each dataset and on the performance of these community detection methods on them.

Political Blog This well-known dataset was recorded by [1] during the 2004 U.S. Presidential Election. The original form is a directed network of hyperlinks between 1490 political blogs. The blogs were manually labeled as either liberal or conservative according to their political leanings. The labels were treated as true community memberships. Following the literature, we removed the direction information and focused on the largest connected component which contains 1222 nodes and 16714 edges. Except for SCORE_{k+1} and OCCAM_{k+1}, all other methods performed comparably on this dataset with Latentnet achieving the smallest misclustered proportion.

Simmons College The Simmons College Facebook network is an undirected graph that contains 1518 nodes and 32988 undirected edges. For comparison purpose, we followed the same pre-processing steps as in [19] by considering the largest connected component of the students with graduation year between 2006 and 2009, which led to a subgraph of 1137 nodes and 24257 edges.

It was observed in [59] that the class year has the highest assortativity values among all available demographic characteristics, and so we treated the class year as the true community label. On this dataset, LSCD_{k+1} and LSCD achieved the two lowest mis-clustered proportions among these methods, with CMM a close third lowest.

An important advantage of model (1) is that it can provide a natural visualization of the network. To illustrate, the left panel of Figure 7 is a 3D visualization of the network with the first three coordinates of the estimated latent variables. From the plot, one can immediately see three big clusters: class year 2006 and 2007 combined (red), class year 2008 (green) and class year 2009 (blue). The right panel zooms into the cluster that includes class year 2006 and 2007 by projecting the the estimated four dimensional latent vectors onto a two dimensional discriminant subspace that was estimated from the fitted latent variables and the clustering results of LSCD. It turned out that class year 2006 and 2007 could also be reasonably distinguished by the latent vectors.

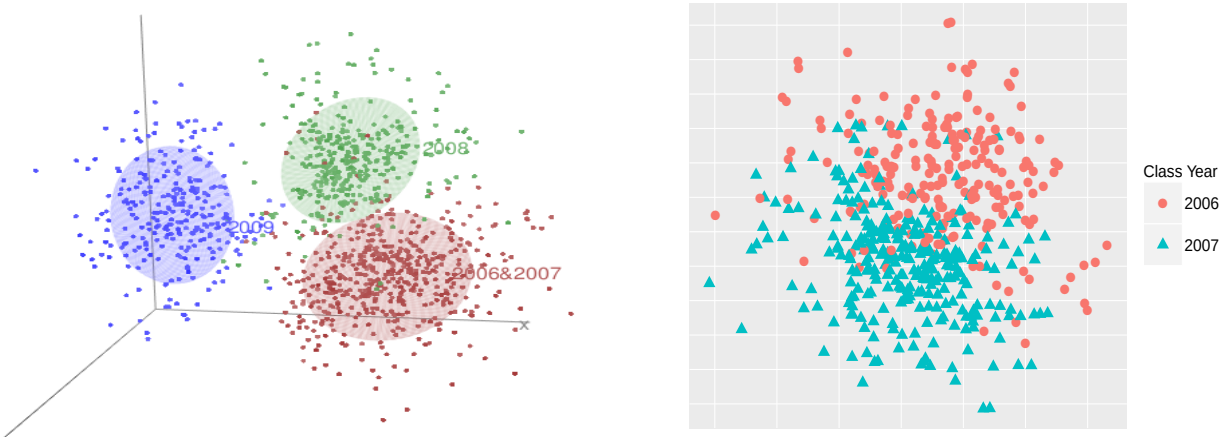

Figure 7: The left panel is a visualization of the network with the first three coordinates of the estimated latent vectors. The right panel is a visualization of students in class year 2006 and 2007 by projecting the four dimensional latent vectors to an estimated two dimensional discriminant subspace.

Caltech Data In contrast to the Simmons College network in which communities are formed according to class years, communities in the Caltech friendship network are formed according to dorms [58, 59]. In particular, students spread across eight different dorms which we treated as true community labels. Following the same pre-processing steps as in [19], we excluded the students whose residence information was missing and considered the largest connected component of the remaining graph, which contained 590 nodes and 12822 undirected edges. This dataset is more challenging than the Simmons College network. Not only the size of the network halves but the number of communities doubles. In some sense, it serves the purpose of testing these methods when the signal is weak. LSCD and LSCD_{k+1} achieved the two highest overall accuracy on this dataset, where LSCD reduced the third best error rate (achieved by CMM) by nearly 15%. See the fourth and last rows of Table 1. Moreover, not taking into account LSCD_{k+1}, SCORE_{k+1} or OCCAM_{k+1}, LSCD achieved the lowest maximum community-wise misclustering error among the other five methods. See Figure 8 for a detailed comparison of community-wise misclustering rates

of the five methods.

It is worth noting that the two spectral methods, SCORE and OCCAM, fell behind on the two Facebook datasets. One possible explanation is that the structures of these Facebook networks are more complex than the political blog network and so DCBM suffers more under-fitting on them. In contrast, the latent space model (1) is more expressive and goes well beyond simple block structure. The Latentnet approach did not perform well on the Facebook datasets, either. One possible reason is the increased numbers of communities compared to the political blog dataset, which substantially increased the difficulty of sampling from posterior distributions.

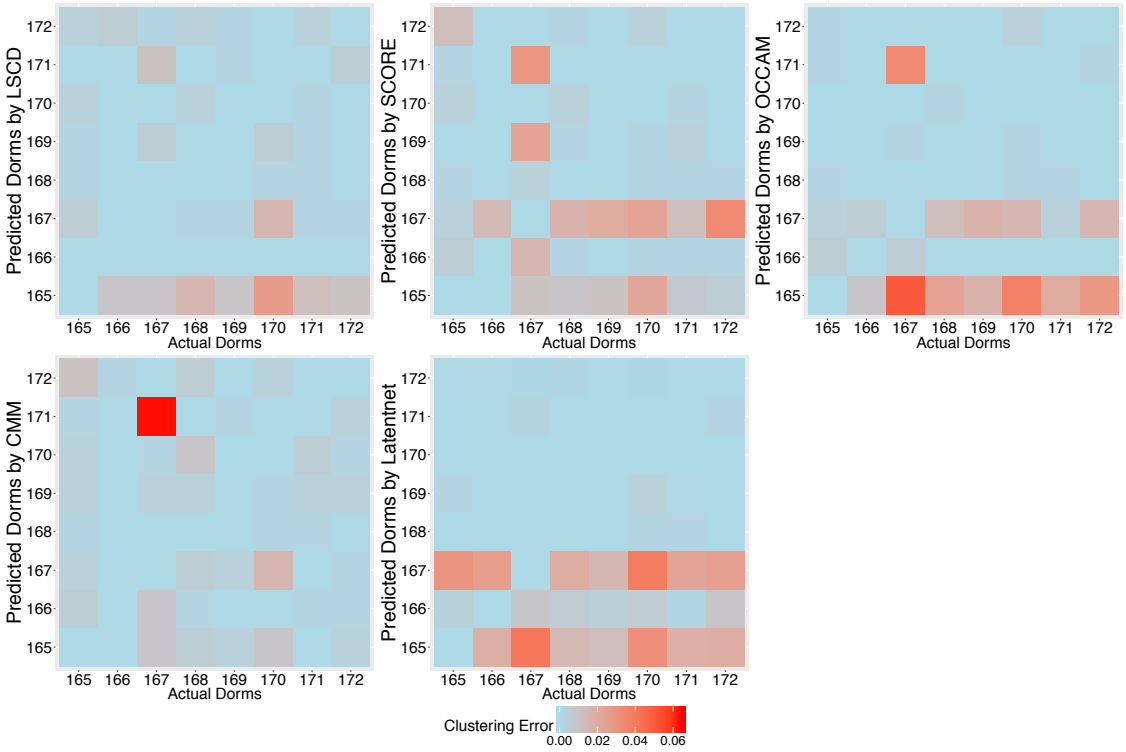

Figure 8: Comparison of community-wise misclustering errors in Caltech friendship network. Top row, left to right: LSCD, SCORE and OCCAM; bottom row, left to right: CMM and Latentnet.

6.2 Community detection with covariate

We now further demonstrate the power of the model and our proposed fitting methods by considering community detection on networks with covariates. Again, we used the LSCD procedure laid out in the previous subsection for community detection.

To this end, we consider a lawyer network dataset which was introduced in [44] that studied the relations among 71 lawyers in a New England law firm. The lawyers were asked to check the names of those who they socialized with outside work, who they knew their family and vice versa. There are also several node attributes contained in the dataset: status (partner or associate), gender, office, years in the firm, age, practice (litigation or corporate), and law school attended, among which status is most assortative. Following [65], we took status as the true community label.

Furthermore, we symmetrized the adjacency matrix, excluded two isolated nodes and finally ended up with 69 lawyers connected by 399 undirected edges.

Visualization and clustering results with and without covariate are shown in Figure 9. On the left panel, as we can see, the latent vectors without adjustment by any covariate worked reasonably well in separating the lawyers of different status and most of the 12 errors (red diamonds) were committed on the boundary. On the right panel, we included a covariate 'practice' into the latent space model: we set $X_{ij} = X_{ji} = 1$ if $i \neq j$ and the ith and the jth lawyers shared the same practice, and $X_{ij} = X_{ji} = 0$ otherwise. Ideally, the influence on the network of being the same type of lawyer should be 'ruled out' this way and the remaining influence on connecting probabilities should mainly be the effect of having different status. In other words, the estimated latent vectors should mainly contain the information of lawyers' status and the effect of lawyers' practice type should be absorbed into the factor βX. The predicted community memberships of lawyers indexed by orange numbers (39, 43, 45, 46, 51, 58) were successfully corrected after introducing this covariate. So the number of mis-clustered nodes was reduced by 50%. We also observed that lawyer 37, though still mis-clustered, was significantly pushed towards the right cluster.

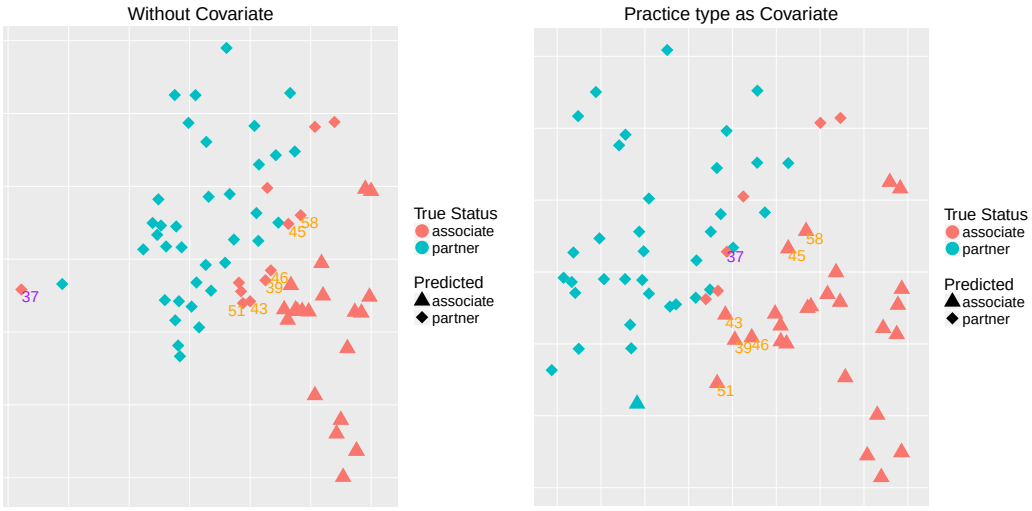

Figure 9: Visualization of the lawyer network using the estimated two dimensional latent vectors. The left panel shows results without including any covariate while the right panel shows results that used practice type information.

6.3 Network assisted learning

In this section, we demonstrate how fitting model (1) can generate new features to be used in machine learning applications when additional network information is available. Consider a network with n nodes and observed adjacency matrix A. Suppose the profile of the nodes is represented by d dimensional features, denoted by $x_1, \cdots, x_n \in \mathbb{R}^d$. Assume each node is associated with a label (or say, variable of interest), denoted by y, either continuous or categorical. Suppose the labels are only observed for a subset of the nodes in the network. Without loss of generality, we assume $y_1, \cdots, y_m$ are observed for some $m < n$. The goal here is to predict the labels $y_{m+1}, \cdots, y_n$ based

on the available information. Without considering the network information, this is the typical setup of supervised learning with labeled training set $(x_1, y_1), \cdots, (x_m, y_m)$ and unlabeled test set $x_{m+1}, \cdots, x_n$. As one way to utilize the network information, we propose to supplement the existing features in the prediction task with the latent vectors estimated by Algorithm 1 (without any edge covariates).

To give a specific example, we considered the Cora dataset [47]. It contains 2708 machine learning papers which were manually classified into 7 categories: Neural Networks, Rule Learning, Reinforcement Learning, Probabilistic Methods, Theory, Genetic Algorithms and Case Based. The dataset also includes the contents of the papers and a citation network, which are represented by a document-word matrix (the vocabulary contains 1433 frequent words) and an adjacency matrix respectively. The task is to predict the category of the papers based on the available information. For demonstration purpose, we only distinguish neural network papers from the other categories, and so the label y is binary.

Let W be the document-word matrix. In the present example, W is of size 2708×1433 (2708 papers and 1433 frequent words). An entry W_{ij} equals 1 if the ith document contains the jth word. Otherwise W_{ij} equals zero. As a common practice in latent semantic analysis, to represent the text information as vectors, we extract leading-d principal component loadings from $WW^\top$ as the features. We chose $d = 100$ by maximizing the prediction accuracy using cross-validation.

However, how to utilize the information contained in the citation network for the desired learning problem is less straightforward. We propose to augment the latent semantic features with the latent vectors estimated from the citation network. Based on the simple intuition that papers in the same category are more likely to cite each other, we expect the latent vectors, as low dimensional summary of the network, to contain information about the paper category. The key message we want to convey here is that with vector representation of the nodes obtained from fitting the latent space model, network information can be incorporated in many supervised and unsupervised learning problems and other exploratory data analysis tasks.

Back to the Cora dataset, for illustration purpose, we fitted standard logistic regressions with the following three sets of features:

1. the leading 100 principal component loadings;
2. estimated degree parameters $\hat{\alpha}_i$ and latent vectors $\hat{z}_i$ obtained from Algorithm 1;
3. the combination of features in 1 and 2.

We considered three different latent space dimensions: $k = 2, 5, 10$. As we can see from Figure 10, the latent vectors contained a considerable amount of predictive power for the label. Adding the latent vectors to the principal components of the word-document matrix could substantially reduce misclassification rate.

7 Discussion

In this section, we discuss a number of related issues and potential problems for future research.

Data-driven choice of latent space dimension For the projected gradient descent method, i.e., Algorithm 1, one needs to specify the latent space dimension k as an input. Although

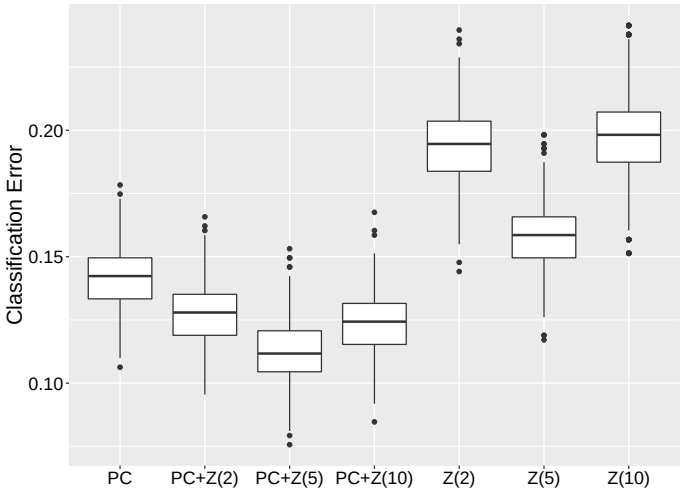

Figure 10: Boxplots of misclassification rates using logistic regression with different feature sets. We randomly split the dataset into training and test sets with size ratio 3:1 for 500 times and computed misclassification errors for each configuration. PC represents the leading 100 principal component loadings of the document-word matrix. $Z(k)$ represents the feature matrix where the ith row is the concatenation of the estimated degree parameter $\widehat{\alpha}_i$ and the estimated latent vector $\widehat{z}_i$ with latent dimension k. PC+$Z(k)$ means the combination of the two sets of features.

Theorem 4.4 suggests that the algorithm could still work reasonably well if the specified latent space dimension is slightly off the target, it is desirable to have a systematic approach to selecting k based on data. One possibility is to inspect the eigenvalues of G^T in Algorithm 2 and set k to be the number of eigenvalues larger than the parameter λ_n used in the algorithm.

Undirected networks with multiple covariates and weighted edges The model (1) and the fitting methods can easily be extended to handle multiple covariates. See also Appendix B. When the number of covariates is fixed, error bounds analogous to those in Section 4 can be expected. We omit the details. Moreover, as pointed out in [29], latent space models for binary networks such as (1) can readily be generalized to weighted networks, i.e., networks with non-binary edges. We refer interested readers to the general recipe spelled out in Section 3.9 of [29]. If the latent variables enter a model for weighted networks in the same way as in model (1), we expect the key ideas behind our proposed fitting methods to continue to work.

Directed networks In many real world networks, edges are directed. Thus, it is a natural next step to generalize model (1) to handle such data. Suppose for any $i \neq j$, $A_{ij} = 1$ if there is an edge pointing from node i to node j, and $A_{ij} = 0$ otherwise. We can consider the following model: for any $i \neq j$,

$$A_{ij} \overset{ind.}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with} \quad \text{logit}(P_{ij}) = \Theta_{ij} = \alpha_i + \gamma_j + \beta X_{ij} + z_i^\top w_j. \tag{25}$$

Here, the α_i's $\in \mathbb{R}$ model degree heterogeneity of outgoing edges while the γ_j's $\in \mathbb{R}$ model heterogeneity of incoming edges. The meaning of β is the same as in model (1). To further

accommodate asymmetry, we associate with each node two latent vectors $z_i, w_i \in \mathbb{R}^k$, where the z_i's are latent variables influencing outgoing edges and the w_i's incoming edges. Such a model has been proposed and used in the study of recommender system [2] and it is also closely connected to the latent eigenmodel proposed in [33] if one further restricts $z_i \in \{w_i, -w_i\}$ for each i. Under this model, the idea behind the convex fitting method in Section 3.1 can be extended. However, it is more challenging to devise a non-convex fitting method with similar theoretical guarantees to what we have in the undirected case. On the other hand, it should be relatively straightforward to further extend the ideas to directed networks with multiple covariates and weighted edges. A recent paper [62] has appeared after the initial posting of the present manuscript, which obtained some interesting results along these directions.

Latent eigenmodel As pointed out by Hoff [33], it is still restrictive to require the latent component G in (2) to be positive semi-definite. In particular, in the same spirit as the latent eigenmodel proposed in [33], it is of great interest to allow the G term in (2) to be any symmetric (as opposed to positive semi-definite) matrix that has a low (effective) rank. In terms of fitting such a model, it is conceivable that the convex approach in (11) will continue to work with the trace penalty term replaced by a generic nuclear norm penalty. On the other hand, to fit the model to large networks, designing non-convex fitting method and establishing their theoretical properties is an interesting topic for further research. Moreover, they may approximate other interesting models, such as mixed membership stochastic blockmodels [3, 5, 36].

8 Proofs of main theorems

We present here the proofs of Theorem 4.3 and Theorem 4.4 since Theorem 4.1 is a corollary of the former and Theorem 4.2 a corollary of the latter. Throughout the proof, let $P = (\sigma(\Theta_{\star,ij}))$ and $P^0 = (P_{ij}\mathbf{1}_{i\neq j})$. Thus, $\mathbb{E}(A) = P^0$. Moreover, for any $\Theta \in \mathbb{R}^{n\times n}$, define

$$h(\Theta) = -\sum_{i,j=1}^{n} \{A_{ij}\Theta_{ij} + \log(1-\sigma(\Theta_{ij}))\}. \tag{26}$$

For conciseness, we denote $\mathcal{F}_g(n, M_1, M_2, X)$ by $\mathcal{F}_g$ throughout the proof. We focus on the case where X is nonzero, and the case of $X = 0$ is simpler.

8.1 Proof of Theorem 4.3

Let $Z_\star \in \mathbb{R}^{n\times k}$ such that $Z_\star Z_\star^\top$ is the best rank k approximation to $G_\star$. For any matrix M, let $\mathrm{col}(M)$ be the subspace spanned by the column vectors of M and $\mathrm{row}(M) = \mathrm{col}(M^\top)$. For any subspace $\mathcal{S}$ of $\mathbb{R}^n$ (or $\mathbb{R}^{n\times n}$), let $\mathcal{S}^\perp$ be its orthogonal complement, and $\mathcal{P}_\mathcal{S}$ the projection operator onto the subspace. The proof relies on the following two lemmas.

Lemma 8.1. *Let $\mathcal{M}_k^\perp = \{M \in \mathbb{R}^{n\times n} : \mathrm{row}(M) \subset \mathrm{col}(Z_\star)^\perp \text{ and } \mathrm{col}(M) \subset \mathrm{col}(Z_\star)^\perp\}$ and $\mathcal{M}_k$ be its orthogonal complement in $\mathbb{R}^{n\times n}$ under trace inner product. If $\lambda_n \geqslant 2\|A-P\|_{\mathrm{op}}$, then for $\bar{G}_k = \mathcal{P}_{\mathcal{M}_k^\perp} G_\star$, we have*

$$\|\Delta_{\widehat{G}}\|_* \leqslant 4\sqrt{2k}\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widehat{\alpha}}1_n^\top\|_{\mathrm{F}} + \frac{2}{\lambda_n}|\langle A-P, \Delta_{\widehat{\beta}}X\rangle| + 4\|\bar{G}_k\|_*.$$

Proof. See Section 8.1.1. □

Lemma 8.2. *For any $k \geqslant 1$ such that Assumption 4.1 holds. Choose $\lambda_n \geqslant \max\{2\|A-P\|_{\mathrm{op}}, 1\}$ and $|\langle A-P, X\rangle| \leqslant \lambda_n\sqrt{k}\|X\|_{\mathrm{F}}$. There exist constants $C>0$ and $0 \leqslant c < 1$ such that*

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \geqslant (1-c)\big(\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2\big) - C\|\bar{G}_k\|_*^2/k, \quad \text{and}$$
$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leqslant (1+c)\big(\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2\big) + C\|\bar{G}_k\|_*^2/k.$$

Proof. See Section 8.1.2. □

Lemma 8.3. *There exist absolute constants c, C such that for any $\Theta \in \mathcal{F}_g$ with probability at least $1-n^{-c}$, the following inequality holds*

$$\|A-P\|_{\mathrm{op}},\ \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} \leqslant C\sqrt{\max\{ne^{-M_2}, \log n\}}.$$

Proof. For any Θ in the parameter space, the off diagonal elements of Θ are uniformly bounded from above by $-M_2$, and so $\max_{i,j} P^0_{ij} \leqslant e^{-M_2}$. Moreover, $\max_i P_{ii} \leqslant 1$ under our assumption. Thus, $\|A-P\|_{\mathrm{op}} \leqslant \|A-P^0\|_{\mathrm{op}} + \|P^0-P\|_{\mathrm{op}} \leqslant \|A-P^0\|_{\mathrm{op}} + 1$. Together with Lemma 8.12, this implies that there exist absolute constants $c_1, C>0$ such that uniformly over the parameter space

$$\mathbb{P}\left(\|A-P\|_{\mathrm{op}} \leqslant C\sqrt{\max\{ne^{-M_2}, \log n\}}\right) \geqslant 1-n^{-c_1}. \tag{27}$$

Since the diagonal entries of X are all zeros, we have $\langle A-P, X\rangle = \langle A-P^0, X\rangle$. Hence, Lemma 8.13 implies that uniformly over the parameter space,

$$\begin{aligned}\mathbb{P}\left(\frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} \leqslant C\sqrt{\max\{ne^{-M_2}, \log n\}}\right) &\geqslant 1 - 3\exp\left(-C^2\max\left\{ne^{-M_2}, \log n\right\}k/8\right) \\ &\geqslant 1 - 3n^{-C^2k/8}.\end{aligned} \tag{28}$$

Combining (27) and (27) finishes the proof. □

Proof of Theorem 4.3 1° We first establish the deterministic bound. Observe that $\widehat{\Theta} = \widehat{\alpha}1_n{}^\top + 1_n\widehat{\alpha}^\top + \widehat{\beta}X + \widehat{G}$ is the optimal solution to (11), and that the true parameter $\Theta_\star = \alpha_\star 1_n{}^\top + 1_n\alpha_\star^\top + \beta_\star X + G_\star$ is feasible. Thus, we have the basic inequality

$$h(\widehat{\Theta}) - h(\Theta_\star) + \lambda_n(\|\widehat{G}\|_* - \|G_\star\|_*) \leqslant 0, \tag{29}$$

where h is defined in (26). For any $\Theta \in \mathcal{F}_g$, $|\Theta_{ij}| \leqslant M_1$ for all i, j and so for $\tau = e^{M_1}/(1+e^{M_1})^2$, the Hessian

$$\nabla^2 h(\Theta) = \mathrm{diag}\big(\mathrm{vec}\big(\sigma(\Theta)\circ(1-\sigma(\Theta))\big)\big) \succeq \tau I_{n^2\times n^2}.$$

For any vector b, $\mathrm{diag}(b)$ is the diagonal matrix with elements of a on its diagonals. For any matrix $B = [b_1, \ldots, b_n] \in \mathbb{R}^{n\times n}$, $\mathrm{vec}(B) \in \mathbb{R}^{n^2}$ is obtained by stacking $b_1, \ldots, b_n$ in order. For any square

matrices A and B, $A \succeq B$ if and only if $A - B$ is positive semi-definite. With the last display, Taylor expansion gives

$$h(\widehat{\Theta}) - h(\Theta_\star) \geqslant \langle \nabla_\Theta h(\Theta_\star), \Delta_{\widehat{\Theta}} \rangle + \frac{\tau}{2} \|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2.$$

On the other hand, triangle inequality implies

$$\lambda_n(\|\widehat{G}\|_* - \|G_\star\|_*) \geqslant -\lambda_n \|\Delta_G\|_*.$$

Together with (29), the last two displays imply

$$\langle \nabla_\Theta h(\Theta_\star), \Delta_{\widehat{\Theta}} \rangle + \frac{\tau}{2} \|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}} - \lambda_n \|\Delta_{\widehat{G}}\|_* \leqslant 0.$$

Triangle inequality further implies

$$\begin{aligned} \frac{\tau}{2}\|\Delta_\Theta\|_F^2 &\leqslant \lambda_n \|\Delta_G\|_* + |\langle \nabla_\Theta h(\Theta_\star), \Delta_{\widehat{G}} + \Delta_{\widehat{\alpha}} 1_n{}^\top + 1_n \Delta_{\widehat{\alpha}}^\top \rangle| + |\Delta_{\widehat{\beta}} \langle \nabla_\Theta h(\Theta_\star), X \rangle| \\ &= \lambda_n \|\Delta_G\|_* + |\langle A - P, \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top \rangle| + |\Delta_{\widehat{\beta}} \langle A - P, X \rangle| \\ &\leqslant \lambda_n \|\Delta_G\|_* + |\langle A - P, \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top \rangle| + \lambda_n \sqrt{k} \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}. \end{aligned} \tag{30}$$

Here the equality is due to the symmetry of $A - P$ and the last inequality is due to the condition imposed on λ_n. We now further upper bound the first two terms on the rightmost side. First, by Lemma 8.1 and the assumption that $|\langle A - P, X \rangle| \leqslant \lambda_n \sqrt{k} \|X\|_{\mathrm{F}}$, we have

$$\|\Delta_G\|_* \leqslant 4\sqrt{2k}\, \|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + 2\sqrt{k}\, \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}} + 4\|\bar{G}_k\|_*. \tag{31}$$

Moreover, Hölder's inequality implies

$$\begin{aligned} |\langle A - P, \Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}} 1_n{}^\top \rangle| &\leqslant \|A - P\|_{\mathrm{op}} (\|\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_*) \\ &= \|A - P\|_{\mathrm{op}} (\|\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}) \\ &\leqslant \frac{\lambda_n}{2} (\|\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}). \end{aligned} \tag{32}$$

Here the equality holds since $\Delta_{\widehat{\alpha}} 1_n{}^\top$ is a rank one matrix. Substituting (31) and (32) into (30), we obtain that

$$\begin{aligned} \frac{\tau}{2}\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 &\leqslant \frac{3\lambda_n}{2} \|\Delta_{\widehat{G}}\|_* + \lambda_n \|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \lambda_n \sqrt{k} \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}} \\ &\leqslant \frac{3\lambda_n}{2} (4\sqrt{2k} \|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + 2\sqrt{k} \|\Delta_{\widehat{\beta}} X\|_F + 4\|\bar{G}_k\|_*) \\ &\qquad + \lambda_n \|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \lambda_n \sqrt{k} \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}} \\ &\leqslant C_1 \lambda_n \big(\sqrt{k}\, (\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widehat{G}}\|_{\mathrm{F}} + \|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \|\Delta_{\widehat{\beta}} X\|_{\mathrm{F}}) + \|\bar{G}_k\|_* \big). \end{aligned}$$

By Lemma 8.2, we can further bound the righthand side as

$$\begin{aligned} \frac{\tau}{2}\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 &\leqslant C_2 \lambda_n \sqrt{k}\, (\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}} + \|\bar{G}_k\|_*/\sqrt{k}) + C_1 \lambda_n \|\bar{G}_k\|_* \\ &\leqslant C_2 \lambda_n \sqrt{k}\, \|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}} + (C_1 + C_2) \lambda_n \|\bar{G}_k\|_*. \end{aligned}$$

Solving the quadratic inequality, we obtain

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leqslant C'\left(\frac{\lambda_n^2 k}{\tau^2}+\frac{\lambda_n\|\bar{G}_k\|_*}{\tau}\right).$$

Note that $\tau \geqslant ce^{-M_1}$ for some positive constant c. Therefore,

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leqslant C\left(e^{2M_1}\lambda_n^2 k+e^{M_1}\lambda_n\|\bar{G}_k\|_*\right).$$

2° We now turn to the probabilistic bound. By Lemma 8.3, there exist constants c_1, C_1 such that for any $\lambda_n \geqslant 2C_1\sqrt{\max\{ne^{-M_2}, \log n\}}$, we have uniformly over the parameter space that

$$\mathbb{P}\left(\lambda_n \geqslant 2\max\left\{\|A-P\|_{\mathrm{op}}, \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}}\right\}\right) \geqslant 1-n^{-c_1}.$$

Denote this event as E. Since the conditions on λ_n in the first part of Theorem 4.3 are satisfied on E, it follows that there exists an absolute constant $C>0$ such that uniformly over the parameter space, with probability at least $1-n^{-c_1}$, $\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leqslant C\phi_n^2$. This completes the proof. □

8.1.1 Proof of Lemma 8.1

By the convexity of $h(\Theta)$,

$$\begin{aligned}
h(\widehat{\Theta})-h(\Theta_\star) &\geqslant \langle \nabla_\Theta h(\Theta_\star), \Delta_{\widehat{\Theta}}\rangle \\
&= -\langle A-P,\, \Delta_{\widehat{G}}+2\Delta_{\widehat{\alpha}}1_n{}^\top+\Delta_{\widehat{\beta}}X\rangle \\
&\geqslant -\|A-P\|_{\mathrm{op}}\left(\|\Delta_{\widehat{G}}\|_*+2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_*\right)-|\langle A-P, \Delta_{\widehat{\beta}}X\rangle| \\
&\geqslant -\frac{\lambda_n}{2}\left(\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_*+\|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_*+2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}\right)-|\langle A-P, \Delta_{\widehat{\beta}}X\rangle|.
\end{aligned}$$

The last inequality holds since $\lambda_n \geqslant 2\|A-P\|_{\mathrm{op}}$ and $\mathcal{P}_{\mathcal{M}_k}+\mathcal{P}_{\mathcal{M}_k^\perp}$ equals identity. On the other hand, by the definition of $\bar{G}_k$,

$$\begin{aligned}
\|\widehat{G}\|_*-\|G_\star\|_* &= \|\mathcal{P}_{\mathcal{M}_k}G_\star+\bar{G}_k+\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}+\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_*-\|\mathcal{P}_{\mathcal{M}_k}G_\star+\bar{G}_k\|_* \\
&\geqslant \|\mathcal{P}_{\mathcal{M}_k}G_\star+\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_*-\|\bar{G}_k\|_*-\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_*-\|\mathcal{P}_{\mathcal{M}_k}G_\star\|_*-\|\bar{G}_k\|_* \\
&= \|\mathcal{P}_{\mathcal{M}_k}G_\star\|_*+\|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_*-2\|\bar{G}_k\|_*-\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_*-\|\mathcal{P}_{\mathcal{M}_k}G_\star\|_* \\
&= \|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_*-\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_*-2\|\bar{G}_k\|_*.
\end{aligned}$$

Here, the second last equality holds since $\mathcal{P}_{\mathcal{M}_k}G_\star$ and $\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}$ have orthogonal column and row spaces. Furthermore, since $\widehat{\Theta}$ is the optimal solution to (11), and $\Theta_\star$ is feasible, the basic inequality and the last two displays imply

$$\begin{aligned}
0 &\geqslant h(\widehat{\Theta})-h(\Theta_\star)+\lambda_n\left(\|\widehat{G}\|_*-\|G_\star\|_*\right) \\
&\geqslant -\frac{\lambda_n}{2}\left(\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_*+\|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_*+2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}\right) \\
&\qquad -|\langle A-P, \Delta_{\widehat{\beta}}X\rangle|+\lambda_n\left(\|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_*-\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_*-2\|\bar{G}_k\|_*\right) \\
&= \frac{\lambda_n}{2}\left(\|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_*-3\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_*-4\|\bar{G}_k\|_*-2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}\right)-|\langle A-P, \Delta_{\widehat{\beta}}X\rangle|.
\end{aligned}$$

Rearranging the terms leads to

$$\|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widehat{G}}\|_* \leqslant 3\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}} + \frac{2}{\lambda_n}|\langle A - P, \Delta_{\widehat{\beta}}X\rangle| + 4\|\bar{G}_k\|_*\,,$$

and triangle inequality further implies

$$\|\Delta_{\widehat{G}}\|_* \leqslant 4\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_* + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}} + \frac{2}{\lambda_n}|\langle A - P, \Delta_{\widehat{\beta}}X\rangle| + 4\|\bar{G}_k\|_*\,.$$

Last but not least, note that the rank of $\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}$ is at most $2k$, and so we complete the proof by further bounding the first term on the righthand side of the last display by $4\sqrt{2k}\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_{\mathrm{F}}$.

8.1.2 Proof of Lemma 8.2

By definition, we have the decomposition

$$\begin{aligned}
\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 &= \|\Delta_{\widehat{G}} + \Delta_{\widehat{\alpha}}1_n{}^\top + 1_n\Delta_{\widehat{\alpha}}^\top + \Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2 \\
&= \|\Delta_{\widehat{G}} + \Delta_{\widehat{\alpha}}1_n{}^\top + 1_n\Delta_{\widehat{\alpha}}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2 + 2\langle\Delta_{\widehat{G}} + \Delta_{\widehat{\alpha}}1_n{}^\top + 1_n\Delta_{\widehat{\alpha}}^\top, \Delta_{\widehat{\beta}}X\rangle \\
&= \|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + 2\ \mathrm{Tr}(\Delta_{\widehat{\alpha}}1_n{}^\top\Delta_{\widehat{\alpha}}1_n{}^\top) + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2 + 2\langle\Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}}1_n{}^\top, \Delta_{\widehat{\beta}}X\rangle\,.
\end{aligned}$$

Here the last equality is due to the symmetry of X and the fact that $\Delta_{\widehat{G}}1_n = 0$. Since $\mathrm{Tr}(\Delta_{\widehat{\alpha}}1_n{}^\top\Delta_{\widehat{\alpha}}1_n{}^\top) = \mathrm{Tr}(1_n{}^\top\Delta_{\widehat{\alpha}}1_n{}^\top\Delta_{\widehat{\alpha}}) = |1_n{}^\top\Delta_{\widehat{\alpha}}|^2 \geqslant 0$, the last display implies

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \geqslant \|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2 + 2\langle\Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}}1_n{}^\top, \Delta_{\widehat{\beta}}X\rangle. \tag{33}$$

Furthermore, we have

$$\begin{aligned}
&|\langle\Delta_{\widehat{G}} + 2\Delta_{\widehat{\alpha}}1_n{}^\top, \Delta_{\widehat{\beta}}X\rangle| \\
&\leqslant \|\Delta_{\widehat{G}}\|_*\|\Delta_{\widehat{\beta}}X\|_{\mathrm{op}} + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_*\|\Delta_{\widehat{\beta}}X\|_{\mathrm{op}} \\
&\leqslant \big(4\sqrt{2k}\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_{\mathrm{F}} + 4\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}} + \frac{2}{\lambda_n}|\langle A - P, \Delta_{\widehat{\beta}}X\rangle| + 4\|\bar{G}_k\|_*\big)\|\Delta_{\widehat{\beta}}X\|_{\mathrm{op}} \\
&\leqslant \big(4\sqrt{2k}\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widehat{G}}\|_{\mathrm{F}} + 4\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}} + 2\sqrt{k}\|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}} + 4\|\bar{G}_k\|_*\big)\frac{\|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}} \\
&\leqslant \frac{C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\big(\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2\big) + \frac{4\|\bar{G}_k\|_*}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}} \\
&\leqslant \frac{C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\big(\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2\big) + \frac{2\|\bar{G}_k\|_*^2}{c_0\,\mathrm{r}_{\mathrm{stable}}(X)} + 2c_0\|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2
\end{aligned}$$

for any constant $c_0 \geqslant 0$. Here, the first inequality holds since the operator norm and the nuclear norm are dual norms under trace inner product. The second inequality is due to Lemma 8.1 and the fact that $\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_* = \|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}$ since $\Delta_{\widehat{\alpha}}1_n{}^\top$ is of rank one. The third inequality is due to the definition of $\mathrm{r}_{\mathrm{stable}}(X)$ and that $|\langle A - P, X\rangle| \leqslant \lambda_n\sqrt{k}\|X\|_{\mathrm{F}}$ by assumption and $\Delta_{\widehat{\beta}}$ is a scalar. The fourth inequality is due to Assumption 4.1 and the last due to $2ab \leqslant a^2 + b^2$ for any $a, b \in \mathbb{R}$.

Substituting these inequalities into (33) leads to

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \geqslant \left(1 - \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + \left(2 - \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \left(1 - \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}} - 4c_0\right)\|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2 - \frac{4\|\bar{G}_k\|_*^2}{c_0\,\mathrm{r}_{\mathrm{stable}}(X)}\,.$$

On the other hand, notice that $\mathrm{Tr}(\Delta_{\widehat{\alpha}}1_n{}^\top\Delta_{\widehat{\alpha}}1_n{}^\top) \leqslant \|\Delta_{\widehat{\alpha}}1_n{}^\top\|_F^2$, we have

$$\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leqslant \left(1 + \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\|\Delta_{\widehat{G}}\|_{\mathrm{F}}^2 + \left(4 + \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\|\Delta_{\widehat{\alpha}}1_n{}^\top\|_{\mathrm{F}}^2 + \left(1 + \frac{2C_0\sqrt{k}}{\sqrt{\mathrm{r}_{\mathrm{stable}}(X)}} + 4c_0\right)\|\Delta_{\widehat{\beta}}X\|_{\mathrm{F}}^2 + \frac{4\|\bar{G}_k\|_*^2}{c_0\,\mathrm{r}_{\mathrm{stable}}(X)}\,.$$

Together with Assumption 4.1, the last two displays complete the proof.

8.2 Proofs of Lemma 4.1 and Theorem 4.4

Again, we directly prove the results under the general model. Recall that $G_\star \approx U_k D_k U_k^\top$ is the top-k eigen-decomposition of $G_\star$, $Z_\star = U_k D_k^{1/2}$, $\bar{G}_k = G_\star - U_k D_k U_k^\top$ and $\Delta_{G^t} = Z^t(Z^t)^\top - Z_\star Z_\star^\top$. For the convenience of analysis, we will instead analyze the following quantity,

$$\widetilde{e}_t = \left\|Z^0\right\|_{\mathrm{op}}^2\|\Delta_{Z^t}\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2.$$

Under Assumption 4.2,

$$\|\Delta_{Z^0}\|_{\mathrm{op}} \leqslant \delta\,\|Z_\star\|_{\mathrm{op}}\,, \quad (1-\delta)e_t \leqslant \widetilde{e}_t \leqslant (1+\delta)e_t. \tag{34}$$

for some sufficiently small constant $\delta \in (0,1)$. The rest of the proof relies on the following lemmas.

Lemma 8.4. *For any $\Theta_\star \in \mathcal{F}_g(n, M_1, M_2, X)$, $\max_{1\leqslant i\leqslant n}\|(Z_\star)_i\|_2^2 \leqslant M_1/3$.*

Proof. By definition, $G_\star - Z_\star Z_\star^\top \in \mathcal{S}_+^n$, which implies, $e_i^\top\left(G_\star - Z_\star Z_\star^\top\right)e_i = G_{ii} - \|(Z_\star)_i\|_2^2 \geqslant 0$, that is $\|(Z_\star)_i\|_2^2 \leqslant G_{ii} \leqslant M_1/3$ for any $1 \leqslant i \leqslant n$. □

Lemma 8.5. *If Assumption 4.1 holds, there exist constants $0 \leqslant c_0 < 1$ and C_0 such that*

$$\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 \geqslant (1-c_0)\left(\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2\right) - C_0\|\bar{G}_k\|_{\mathrm{F}}^2,$$
$$\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 \leqslant (1+c_0)\left(\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2\right) + C_0\|\bar{G}_k\|_{\mathrm{F}}^2.$$

Proof. See Section 8.2.1. □

Lemma 8.6. *Under Assumption 4.1, let $\zeta_n = \max\{2\,\|A-P\|_{\mathrm{op}}\,,\ |\langle A-P, X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$, if $\|\Delta_{Z^t}\|_{\mathrm{F}} \leqslant c_0 e^{-M_1}\|Z_\star\|_{\mathrm{op}}/\kappa_{Z_\star}^2$ and $\|Z_\star\|_{\mathrm{op}}^2 \geqslant C_0 e^{M_1}\kappa_{Z_\star}^2\zeta_n^2$ for sufficiently small constant c_0 and sufficiently large constant C_0, there exist a constant c such that, for any $\eta \leqslant c$, there exist positive constants ρ and C,*

$$\widetilde{e}_{t+1} \leqslant \left(1 - \frac{\eta}{e^{M_1}\kappa^2}\rho\right)\widetilde{e}_t + \eta C\left(\|\bar{G}_k\|_{\mathrm{F}}^2 + e^{M_1}\zeta_n^2 k\right).$$

Proof. See Section 8.2.2. □

Lemma 8.7. *Under Assumption 4.1, let* $\zeta_n = \max\{2\|A-P\|_{\mathrm{op}},\ |\langle A-P, X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$, *if* $\|Z_\star\|_{\mathrm{op}}^2 \geqslant C_1\kappa_{Z_\star}^2\zeta_n^2 e^{M_1}\max\left\{\sqrt{\eta\|\bar{G}_k\|_{\mathrm{F}}^2/\zeta_n^2}, \sqrt{\eta k e^{M_1}}, 1\right\}$ *for a sufficiently large constant* C_1 *and* $\widetilde{e}_0 \leqslant c_0^2 e^{-2M_1}\|Z_\star\|_{\mathrm{op}}^4/4\kappa_{Z_\star}^4$, *then for all* $t \geqslant 0$,

$$\|\Delta_{Z^t}\|_{\mathrm{F}} \leqslant \frac{c_0}{e^{M_1}\kappa_{Z_\star}^2}\|Z_\star\|_{\mathrm{op}}.$$

Proof. See Section 8.2.3. □

Proof of Lemma 4.1 By Lemma 8.5, notice that $\bar{G}_k = 0$ under the inner product model,

$$\begin{aligned}\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 &\geqslant (1-c_0)\left(\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2\right),\\ \|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 &\leqslant (1+c_0)\left(\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2\right).\end{aligned} \tag{35}$$

By Lemma 8.9,

$$\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 \geqslant 2(\sqrt{2}-1)\kappa_{Z_\star}^{-2}\|Z_\star\|_{\mathrm{op}}^2\|\Delta_{Z^t}\|_{\mathrm{F}}^2$$

which implies,

$$e_t \leqslant \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)}\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2 \leqslant \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)(1-c_0)}\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2.$$

Similarly, by Lemma 8.10, when $\mathrm{dist}(Z^t, Z_\star) \leqslant c\|Z_\star\|_{\mathrm{op}}$,

$$\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 \leqslant (2+c)^2\|Z_\star\|_{\mathrm{op}}^2\|\Delta_{Z^t}\|_{\mathrm{F}}^2,$$

and this implies,

$$\begin{aligned}e_t &\geqslant \frac{1}{(2+c)^2}\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2\\ &\geqslant \frac{1}{(2+c)^2(1+c_0)}\|Z_\star\|_{\mathrm{op}}^2\|\Delta_{Z^t}\|_{\mathrm{F}}^2.\end{aligned}$$

□

Proof of Theorem 4.4 Consider the deterministic bound first. By Lemma 8.7, for all $t \geqslant 0$,

$$\|\Delta_{Z^t}\|_{\mathrm{F}} \leqslant \frac{c_0}{e^{M_1}\kappa_{Z_\star}^2}\|Z_\star\|_{\mathrm{op}}.$$

Then apply Lemma 8.6, there exists positive constants ρ and M such that for all $t \geqslant 0$,

$$\widetilde{e}_{t+1} \leqslant \left(1 - \frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)\widetilde{e}_t + \eta C\left(\|\bar{G}_k\|_{\mathrm{F}}^2 + e^{M_1}\zeta_n^2 k\right).$$

Therefore,

$$\begin{aligned}\widetilde{e}_t &\leqslant \left(1-\frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t \widetilde{e}_0 + \sum_{i=0}^{t}\eta C\left(\|\bar{G}_k\|_{\mathrm{F}}^2 + e^{M_1}\zeta_n^2 k\right)\left(1-\frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^i \\ &\leqslant \left(1-\frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t \widetilde{e}_0 + \frac{C\kappa^2}{\rho}\left(e^{2M_1}\zeta_n^2 k + e^{M_1}\|\bar{G}_k\|_{\mathrm{F}}^2\right).\end{aligned}$$

Notice that $0.9e_t \leqslant \widetilde{e}_t \leqslant 1.1e_t$,

$$e_t \leqslant 2\left(1-\frac{\eta}{e^{M_1}\kappa_{Z_\star}^2}\rho\right)^t e_0 + \frac{2C\kappa^2}{\rho}\left(e^{2M_1}\zeta_n^2 k + e^{M_1}\|\bar{G}_k\|_{\mathrm{F}}^2\right).$$

Given the last display, the proof of the probabilistic bound is nearly the same as that of the counterpart in Theorem 4.3 and we leave out the details. □

8.2.1 Proof of Lemma 8.5

By definition,

$$\begin{aligned}\|\Delta_{G^t}\|_{\mathrm{F}}^2 &= \|Z^t(Z^t)^\top - Z_\star Z_\star^\top - \bar{G}_k\|_{\mathrm{F}}^2 \\ &\geqslant \|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + \|\bar{G}_k\|_{\mathrm{F}}^2 - 2|\big\langle Z^t(Z^t)^\top - Z_\star Z_\star^\top, \bar{G}_k\big\rangle| \\ &\geqslant \|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + \|\bar{G}_k\|_{\mathrm{F}}^2 - 2\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}\|\bar{G}_k\|_{\mathrm{F}} \\ &\geqslant \|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + \|\bar{G}_k\|_{\mathrm{F}}^2 - c_1\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 - c_1^{-1}\|\bar{G}_k\|_{\mathrm{F}}^2 \\ &\geqslant (1-c_1)\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 - (c_1^{-1}-1)\|\bar{G}_k\|_{\mathrm{F}}^2\end{aligned}$$

where the second last inequality comes from $a^2+b^2 \geqslant 2ab$ and holds for any $c_1 \geqslant 0$. Similarly, it could be shown that

$$\|\Delta_{G^t}\|_{\mathrm{F}}^2 \leqslant (1+c_1)\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + (1+c_1^{-1})\|\bar{G}_k\|_{\mathrm{F}}^2.$$

Expanding the term $\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2$, we obtain

$$\begin{aligned}\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 &= \|\Delta_{G^t} + \Delta_{\alpha^t}1_n{}^\top + 1_n\Delta_{\alpha^t}^\top + \Delta_{\beta^t}X\|_{\mathrm{F}}^2 \\ &= \|\Delta_{G^t} + \Delta_{\alpha^t}1_n{}^\top + 1_n\Delta_{\alpha^t}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2 + 2\big\langle\Delta_{G^t} + \Delta_{\alpha^t}1_n{}^\top + 1_n\Delta_{\alpha^t}^\top, \Delta_{\beta^t}X\big\rangle \\ &= \|\Delta_{G^t}\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + 2\,\mathrm{Tr}(\Delta_{\alpha^t}1_n{}^\top\Delta_{\alpha^t}1_n{}^\top) + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2 \\ &\quad + 2\big\langle\Delta_{G^t} + 2\Delta_{\alpha^t}1_n{}^\top, \Delta_{\beta^t}X\big\rangle,\end{aligned}$$

where the last equality is due to the symmetry of X. Notice that $\mathrm{Tr}(\Delta_{\widehat{\alpha}}1_n{}^\top\Delta_{\widehat{\alpha}}1_n{}^\top) = \mathrm{Tr}(1_n{}^\top\Delta_{\widehat{\alpha}}1_n{}^\top\Delta_{\widehat{\alpha}}) = |1_n{}^\top\Delta_{\widehat{\alpha}}|^2 \geqslant 0$,

$$\begin{aligned}\|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 \geqslant{}& (1-c_1)\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 - (c_1^{-1}-1)\|\bar{G}_k\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2 \\ &+ 2\big\langle Z^t(Z^t)^\top - Z_\star Z_\star^\top + 2\Delta_{\alpha^t}1_n{}^\top, \Delta_{\beta^t}X\big\rangle - 2\big\langle\bar{G}_k, \Delta_{\beta^t}X\big\rangle.\end{aligned} \tag{36}$$

By Hölder's inequality,

$$\begin{aligned}
|\big\langle Z^t(Z^t)^\top - Z_\star Z_\star^\top + 2\Delta_{\alpha^t} 1_n{}^\top, \Delta_{\beta^t} X\big\rangle| &\leqslant \left(\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_* + 2\|\Delta_{\alpha^t} 1_n{}^\top\|_*\right) \left\|\Delta_{\beta^t} X\right\|_{\mathrm{op}} \\
&\leqslant \left(\sqrt{2k}\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}} + 2\|\Delta_{\alpha^t} 1_n{}^\top\|_{\mathrm{F}}\right) \left\|\Delta_{\beta^t} X\right\|_{\mathrm{op}} \\
&\leqslant \left(\sqrt{2k}\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}} + 2\|\Delta_{\alpha^t} 1_n{}^\top\|_{\mathrm{F}}\right) \|\Delta_{\beta^t} X\|_{\mathrm{F}}/\sqrt{\mathrm{r}_{\mathrm{stable}}(X)} \\
&\leqslant C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} \left(\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + \|\Delta_{\alpha^t} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t} X\|_{\mathrm{F}}^2\right),
\end{aligned}$$

and for any $c > 0$,

$$|\big\langle \bar{G}_k, \Delta_{\beta^t} X\big\rangle| \leqslant \|\bar{G}_k\|_{\mathrm{F}} \|\Delta_{\beta^t} X\|_{\mathrm{F}} \leqslant c\|\Delta_{\beta^t} X\|_{\mathrm{F}}^2 + \frac{1}{4c}\|\bar{G}_k\|_{\mathrm{F}}^2.$$

Substitute these inequalities into (36),

$$\begin{aligned}
\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \geqslant &\left(1 - 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} - c_1\right)\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + \left(2 - 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\|\Delta_{\alpha^t} 1_n{}^\top\|_{\mathrm{F}}^2 \\
&+ \left(1 - 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} - 2c\right)\|\Delta_{\beta^t} X\|_{\mathrm{F}}^2 - (c_1^{-1} + 1/2c)\|\bar{G}_k\|_{\mathrm{F}}^2.
\end{aligned}$$

On the other hand, notice that $\mathrm{Tr}(\Delta_{\alpha^t} 1_n{}^\top \Delta_{\alpha^t} 1_n{}^\top) \leqslant \|\Delta_{\alpha^t} 1_n\|_F^2$, we have

$$\begin{aligned}
\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \leqslant &\left(1 + 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} + c_1\right)\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + \left(2 + 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}}\right)\|\Delta_{\widehat{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 \\
&+ \left(1 + 2C_1\sqrt{\frac{k}{\mathrm{r}_{\mathrm{stable}}(X)}} + 2c\right)\|\Delta_{\beta^t} X\|_{\mathrm{F}}^2 + (c_1^{-1} + 1/2c)\|\bar{G}_k\|_{\mathrm{F}}^2.
\end{aligned}$$

This completes the proof.

8.2.2 Proof of Lemma 8.6

Let $\Theta^t = \alpha^t 1_n{}^\top + 1_n(\alpha^t)^\top + \beta^t X + Z^t(Z^t)^\top$, $R^t = \underset{R\in\mathbb{R}^{r\times r}, RR^\top = I_r}{\arg\min} \|Z^t - Z_\star R\|_{\mathrm{F}}$, $\widetilde{R}^t = \underset{R\in\mathbb{R}^{r\times r}, RR^\top = I_r}{\arg\min} \|\widetilde{Z}_t - Z_\star R\|_{\mathrm{F}}$ and $\Delta_{Z^t} = Z^t - Z_\star R^t$, then

$$\|Z^{t+1} - Z_\star R^{t+1}\|_{\mathrm{F}}^2 \leqslant \|Z^{t+1} - Z_\star \widetilde{R}^{t+1}\|_{\mathrm{F}}^2 \leqslant \|\widetilde{Z}_{t+1} - Z_\star \widetilde{R}^{t+1}\|_{\mathrm{F}}^2 \leqslant \|\widetilde{Z}_{t+1} - Z_\star R^t\|_{\mathrm{F}}^2.$$

The first and the last inequalities are due to the definition of R^{t+1} and $\widetilde{R}^{t+1}$, and the second inequality is due to the projection step. Plugging in the definition of $\widetilde{Z}^{t+1}$, we obtain

$$\begin{aligned}
\|Z^{t+1} - Z_\star R^{t+1}\|_{\mathrm{F}}^2 &\leqslant \|Z^t - Z_\star R^t\|_{\mathrm{F}}^2 + \eta_Z^2\|\nabla h(\Theta^t) Z^t\|_{\mathrm{F}}^2 - 2\eta_Z\big\langle \nabla h(\Theta^t) Z^t, Z^t - Z_\star R^t\big\rangle \\
&= \|Z^t - Z_\star R^t\|_{\mathrm{F}}^2 + \eta_Z^2\|\nabla h(\Theta^t) Z^t\|_{\mathrm{F}}^2 - 2\eta_Z\big\langle \nabla h(\Theta^t), (Z^t - Z_\star R^t)(Z^t)^\top\big\rangle.
\end{aligned}$$

Note that

$$Z^t(Z^t)^\top - Z_\star R^t(Z^t)^\top = \frac{1}{2}(Z^t(Z^t)^\top - Z_\star Z_\star^\top) + \frac{1}{2}(Z^t(Z^t)^\top + Z_\star Z_\star^\top) - Z_\star R(Z^t)^\top.$$

Also due to the symmetry of $\nabla h(\Theta^t)$,

$$\big\langle \nabla h(\Theta^t), \frac{1}{2}(Z^t(Z^t)^\top + Z_\star Z_\star^\top) - Z_\star R(Z^t)^\top \big\rangle = \frac{1}{2}\big\langle \nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle.$$

Therefore, combine the above three equations,

$$\begin{aligned} \|Z^{t+1} - Z_\star R^{t+1}\|_{\mathrm{F}}^2 \leqslant \|Z^t - Z_\star R^t\|_{\mathrm{F}}^2 + \eta_Z^2\|\nabla h(\Theta^t)Z^t\|_{\mathrm{F}}^2 - \eta_Z\big\langle \nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle \\ - \eta_Z\big\langle \nabla h(\Theta^t), (Z^t(Z^t)^\top - Z_\star Z_\star^\top)\big\rangle. \end{aligned} \tag{37}$$

By similar and slightly simpler arguments, we also obtain

$$\begin{aligned} \|\alpha^{t+1} - \alpha_\star\|^2 &\leqslant \|\widetilde{\alpha}_{t+1} - \alpha_\star\|^2 \\ &= \|\alpha^t - \alpha_\star\|^2 + \eta_\alpha^2\|\nabla h(\Theta^t)1_n\|_{\mathrm{F}}^2 - 2\eta_\alpha\big\langle \nabla h(\Theta^t)1_n, \alpha^t - \alpha_\star\big\rangle. \end{aligned} \tag{38}$$

$$\begin{aligned} \|\beta^{t+1} - \beta_\star\|^2 &\leqslant \|\widetilde{\beta}_{t+1} - \beta_\star\|^2 \\ &= \|\beta^t - \beta_\star\|^2 + \eta_\beta^2\big\langle \nabla h(\Theta^t), X\big\rangle^2 - 2\eta_\beta\big\langle \nabla h(\Theta^t), (\beta^t - \beta_\star)X\big\rangle. \end{aligned} \tag{39}$$

For $h(\Theta)$ in (26), define

$$H(\Theta) = \mathbb{E}_{\Theta_\star}[h(\Theta)] - \sum_{i=1}^n \Theta_{ii}\sigma(\Theta_{\star,ii}).$$

Then it is straightforward to verify that $\nabla H(\Theta) = \sigma(\Theta) - \sigma(\Theta_\star)$ and so $\nabla H(\Theta_\star) = 0$. With $\eta_Z = \eta/\|Z^0\|_{\mathrm{op}}^2, \eta_\alpha = \eta/2n, \eta_\beta = \eta/2\|X\|_{\mathrm{F}}^2$, the weighted sum $\|Z^0\|_{\mathrm{op}}^2 \times (37) + 2n \times (38) + \|X\|_{\mathrm{F}}^2 \times (39)$ is equivalent to

$$\begin{aligned} \widetilde{e}_{t+1} &\leqslant \widetilde{e}_t - \eta\big\langle \nabla h(\Theta^t), Z^t(Z^t)^\top - Z_\star Z_\star^\top + 2(\alpha^t - \alpha_\star)1_n{}^\top + (\beta^t - \beta_\star)X + \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle \\ &\quad + \Big(\|Z^0\|_{\mathrm{op}}^2\,\eta_Z^2\|\nabla h(\Theta^t)Z^t\|_{\mathrm{F}}^2 + 2n\eta_\alpha^2\|\nabla h(\Theta^t)1_n\|_{\mathrm{F}}^2 + \|X\|_{\mathrm{F}}^2\eta_\beta^2\big\langle \nabla h(\Theta^t), X\big\rangle^2\Big) \\ &\leqslant \widetilde{e}_t - \eta\big\langle \nabla h(\Theta^t), \Delta_{\bar{\Theta}^t}\big\rangle - \eta\big\langle \nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle \\ &\quad + \left(\frac{\eta^2}{\|Z^0\|_{\mathrm{op}}^2}\|\nabla h(\Theta^t)Z^t\|_{\mathrm{F}}^2 + \frac{\eta^2}{2n}\|\nabla h(\Theta^t)1_n\|_{\mathrm{F}}^2 + \frac{\eta^2}{4\|X\|_{\mathrm{F}}^2}\big\langle \nabla h(\Theta^t), X\big\rangle^2\right), \end{aligned}$$

where $\Delta_{\bar{\Theta}^t} = Z^t(Z^t)^\top - Z_\star Z_\star^\top + \Delta_{\alpha^t}1_n{}^\top + 1_n(\Delta_{\alpha^t})^\top + \Delta_{\beta^t}X = \Delta_{\Theta^t} - \bar{G}_k$. Then, simple algebra further leads to

$$\begin{aligned} \widetilde{e}_{t+1} &\leqslant \widetilde{e}_t - \eta\big\langle \nabla h(\Theta^t) - \nabla H(\Theta^t), \Delta_{\bar{\Theta}^t}\big\rangle - \eta\big\langle \nabla H(\Theta^t), \Delta_{\Theta^t}\big\rangle - \eta\big\langle \nabla H(\Theta^t), \bar{G}_k\big\rangle - \eta\big\langle \nabla h(\Theta^t), \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle \\ &\quad + \left(\frac{\eta^2}{\|Z^0\|_{\mathrm{op}}^2}\|\nabla h(\Theta^t)Z^t\|_{\mathrm{F}}^2 + \frac{\eta^2}{2n}\|\nabla h(\Theta^t)1_n\|_{\mathrm{F}}^2 + \frac{\eta^2}{4\|X\|_{\mathrm{F}}^2}\big\langle \nabla h(\Theta^t), X\big\rangle^2\right) \\ &\leqslant \widetilde{e}_t - \eta\big\langle \nabla H(\Theta^t), \Delta_{\Theta^t}\big\rangle + \eta\big|\big\langle \nabla h(\Theta^t) - \nabla H(\Theta^t), \Delta_{\bar{\Theta}^t}\big\rangle\big| + \eta\big|\big\langle \nabla h(\Theta), \Delta_{Z^t}\Delta_{Z^t}^\top \big\rangle\big| \\ &\quad + \eta\big|\big\langle \nabla H(\Theta^t), \bar{G}_k\big\rangle\big| + \eta^2\left(\frac{1}{\|Z^0\|_{\mathrm{op}}^2}\|\nabla h(\Theta^t)Z^t\|_{\mathrm{F}}^2 + \frac{1}{2n}\|\nabla h(\Theta^t)1_n\|_{\mathrm{F}}^2 + \frac{1}{4\|X\|_{\mathrm{F}}^2}\big\langle \nabla h(\Theta^t), X\big\rangle^2\right) \\ &= \widetilde{e}_t - \eta D_1 + \eta D_2 + \eta D_3 + \eta D_4 + \eta^2 D_5. \end{aligned} \tag{40}$$

In what follows, we control Note that for any $\Theta \in \mathcal{F}_g$,

$$\frac{1}{4} I_{n^2 \times n^2} \succeq \nabla^2 H(\Theta) = \operatorname{diag}\Big(\operatorname{vec}\big(\sigma(\Theta) \circ (1 - \sigma(\Theta))\big)\Big) \succeq \tau I_{n^2 \times n^2}$$

where $\tau = e^{M_1}/(1+e^{M_1})^2 \asymp e^{-M_1}$. Hence $H(\cdot)$ is τ-strongly convex and $\frac{1}{4}$-smooth. Further notice that $\nabla H(\Theta_\star) = 0$, then by Lemma 8.11,

$$D_1 = \big\langle \nabla H(\Theta^t), \Delta_{\Theta^t} \big\rangle \geqslant \frac{\tau/4}{\tau + 1/4} \|\Delta_{\Theta^t}\|_{\mathrm{F}}^2 + \frac{1}{\tau + 1/4} \|\sigma(\Theta^t) - \sigma(\Theta_\star)\|_{\mathrm{F}}^2.$$

By triangle inequality,

$$D_2 \leqslant |\big\langle \sigma(\Theta_\star) - A, Z^t (Z^t)^\top - Z_\star Z_\star^\top \big\rangle| + 2|\big\langle \sigma(\Theta_\star) - A, \Delta_{\alpha^t} 1_n{}^\top \big\rangle| + |\big\langle \sigma(\Theta_\star) - A, \Delta_{\beta^t} X \big\rangle|.$$

Recall that $\zeta_n = \max\{2\,\|A - P\|_{\mathrm{op}},\ |\langle A - P, X/\|X\|_{\mathrm{F}}\rangle|/\sqrt{k},\ 1\}$, and so

$$D_2 \leqslant \frac{\zeta_n}{2} \|Z^t (Z^t)^\top - Z_\star Z_\star^\top\|_* + \zeta_n \|\Delta_{\alpha^t} 1_n{}^\top\|_* + \zeta_n \sqrt{k} \|\Delta_{\beta^t} X\|_{\mathrm{F}}.$$

Notice that $Z^t (Z^t)^\top - Z_\star Z_\star^\top$ has rank at most $2k$,

$$D_2 \leqslant \frac{\zeta_n \sqrt{2k}}{2} \|Z^t (Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}} + \zeta_n \|\Delta_{\alpha^t} 1_n{}^\top\|_{\mathrm{F}} + \zeta_n \sqrt{k} \|\Delta_{\beta^t} X\|_{\mathrm{F}}.$$

Further by Cauchy-Schwarz inequality, there exists constant C_2 such that for any positive constant c_2 which we will specify later,

$$D_2 \leqslant c_2 \left(\|Z^t (Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t} X\|_{\mathrm{F}}^2 \right) + \frac{C_2}{4c_2} \zeta_n^2 k.$$

By Lemma 8.5, there exist constants c_1, C_1 such that

$$\begin{aligned} D_1 - D_2 \geqslant & \left(\frac{(1-c_1)\tau}{4\tau + 1} - c_2 \right) \left(\|Z^t (Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t} 1_n{}^\top\|_F^2 + \|\Delta_{\beta^t} X\|_{\mathrm{F}}^2 \right) \\ & + \frac{1}{\tau + 1/4} \|\sigma(\Theta^t) - \sigma(\Theta_\star)\|_{\mathrm{F}}^2 - C_1 \|\bar{G}_k\|_{\mathrm{F}}^2 - \frac{C_2}{4c_2} \zeta_n^2 k. \end{aligned} \tag{41}$$

By Lemma 8.9,

$$D_1 - D_2 \geqslant \frac{2(\sqrt{2}-1)}{\kappa^2} \left(\frac{(1-c_1)\tau}{4\tau + 1} - c_2 \right) e_t + \frac{1}{\tau + 1/4} \|\sigma(\Theta^t) - \sigma(\Theta_\star)\|_{\mathrm{F}}^2 - C_1 \|\bar{G}_k\|_{\mathrm{F}}^2 - \frac{C_2}{4c_2} \zeta_n^2 k.$$

To bound D_3, notice that $\Delta_{Z^t} \Delta_{Z^t}^\top$ is a positive semi-definite matrix,

$$\begin{aligned} D_3 &\leqslant |\big\langle \nabla h(\Theta^t), \Delta_{Z^t} \Delta_{Z^t}^\top \big\rangle| \leqslant \big\|\nabla h(\Theta^t)\big\|_{\mathrm{op}} \big\|\Delta_{Z^t} \Delta_{Z^t}^\top\big\|_* \\ &= \big\|\nabla h(\Theta^t)\big\|_{\mathrm{op}} \operatorname{Tr}(\Delta_{Z^t} \Delta_{Z^t}^\top) \leqslant \big\|\nabla h(\Theta^t)\big\|_{\mathrm{op}} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 \\ &= \big\|\nabla h(\Theta^t) - \nabla H(\Theta^t) + \nabla H(\Theta^t)\big\|_{\mathrm{op}} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 \\ &\leqslant \big\|\nabla h(\Theta^t) - \nabla H(\Theta^t)\big\|_{\mathrm{op}} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 + \big\|\nabla H(\Theta^t)\big\|_{\mathrm{op}} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 \\ &= \|\sigma(\Theta_\star) - A\|_{\mathrm{op}} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 + \big\|\sigma(\Theta^t) - \sigma(\Theta_\star)\big\|_{\mathrm{op}} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 \\ &\leqslant \frac{\zeta_n}{2} \|\Delta_{Z^t}\|_{\mathrm{F}}^2 + \|\sigma(\Theta^t) - \sigma(\Theta_\star)\|_{\mathrm{F}} \|\Delta_{Z^t}\|_{\mathrm{F}}^2. \end{aligned}$$

By the assumption that $\|\Delta_{Z^t}\|_{\mathrm{F}} \leqslant \frac{c_0}{e^{M_1}\kappa^2}\|Z_\star\|_{\mathrm{op}}$,

$$\begin{aligned}\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}\|\Delta_{Z^t}\|_{\mathrm{F}}^2 &\leqslant \frac{c_0}{e^{M_1}\kappa^2}\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}\|\Delta_{Z^t}\|_{\mathrm{F}}\|Z_\star\|_{\mathrm{op}}\\ &\leqslant c_3\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}^2+\frac{c_0}{4c_3e^{M_1}\kappa^2}\|\Delta_{Z^t}\|_{\mathrm{F}}^2\|Z_\star\|_{\mathrm{op}}^2.\end{aligned}$$

for any constant c_3 to be specified later. Then

$$D_3 \leqslant \left(\frac{\zeta_n}{2\|Z_\star\|_{\mathrm{op}}^2}+\frac{c_0}{4c_3e^{M_1}\kappa^2}\right)e_t + c_3\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}^2.$$

By the assumption that $\|Z_\star\|_{\mathrm{op}}^2 \geqslant C_0\kappa^2\zeta_n e^{M_1}$ for sufficiently large constant C_0,

$$D_3 \leqslant \left(\frac{1}{2C_0e^{M_1}\kappa^2}+\frac{c_0}{4c_3e^{M_1}\kappa^2}\right)e_t + c_3\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}^2. \tag{42}$$

For D_4 simple algebra leads to

$$\begin{aligned}D_4 &= |\langle \nabla H(\Theta^t), \bar{G}_k\rangle| = |\langle \sigma(\Theta^t)-\sigma(\Theta_\star), \bar{G}_k\rangle| \leqslant \|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}\|\bar{G}_k\|_{\mathrm{F}}\\ &\leqslant c_4\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}^2+\frac{1}{4c_4}\|\bar{G}_k\|_{\mathrm{F}}^2\end{aligned} \tag{43}$$

for any constant c_4 to be specified later.

We now turn to bounding D_5. To this end, we upper bound its three terms separately as follows. First,

$$\begin{aligned}\|\nabla h(\Theta^t)Z^t\|_{\mathrm{F}}^2 &= \|\left(\nabla h(\Theta^t)-\nabla H(\Theta^t)\right)Z^t+\nabla H(\Theta^t)Z^t\|_{\mathrm{F}}^2\\ &\leqslant 2(\|(\nabla h(\Theta^t)-\nabla H(\Theta^t))Z^t\|_{\mathrm{F}}^2+\|\nabla H(\Theta^t)Z^t\|_{\mathrm{F}}^2)\\ &\leqslant 2(\|(\sigma(\Theta_\star)-A)Z^t\|_{\mathrm{F}}^2+\|(\sigma(\Theta^t)-\sigma(\Theta_\star))Z^t\|_{\mathrm{F}}^2)\\ &\leqslant 2(\|\sigma(\Theta_\star)-A\|_{\mathrm{op}}^2\|Z^t\|_{\mathrm{F}}^2+\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}^2\|Z^t\|_{\mathrm{op}}^2)\\ &\leqslant 2\Big(\frac{\zeta_n^2}{4}\|Z^t\|_{\mathrm{F}}^2+\|Z^t\|_{\mathrm{op}}^2\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}^2\Big).\end{aligned}$$

Next,

$$\begin{aligned}\|\nabla h(\Theta^t)1_n\|^2 &= \big\|\left(\nabla h(\Theta^t)-\nabla H(\Theta^t)\right)1_n+\nabla H(\Theta^t)1_n\big\|^2\\ &\leqslant 2\Big(\big\|\left(\nabla h(\Theta^t)-\nabla H(\Theta^t)\right)1_n\big\|^2+\|\nabla H(\Theta^t)1_n\|^2\Big)\\ &\leqslant 2\Big(\|(\sigma(\Theta_\star)-A)1_n\|^2+\big\|(\sigma(\Theta^t)-\sigma(\Theta_\star))1_n\big\|^2\Big)\\ &\leqslant 2n\Big(\frac{\zeta_n^2}{4}+\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}^2\Big).\end{aligned}$$

Furthermore,

$$\begin{aligned}\big\langle \nabla H(\Theta^t), X\big\rangle^2 &= \Big(\big\langle \nabla h(\Theta^t)-\nabla H(\Theta^t), X\big\rangle+\big\langle \nabla H(\Theta^t), X\big\rangle\Big)^2\\ &\leqslant 2\Big(\big\langle \sigma(\Theta_\star)-A, X\big\rangle^2+\big\langle \sigma(\Theta^t)-\sigma(\Theta_\star), X\big\rangle^2\Big)\\ &\leqslant 2\Big(\zeta_n^2k\|X\|_{\mathrm{F}}^2+\|\sigma(\Theta^t)-\sigma(\Theta_\star)\|_{\mathrm{F}}^2\|X\|_{\mathrm{F}}^2\Big).\end{aligned}$$

When $\operatorname{dist}(Z^t, Z_\star) \leqslant c \left\|Z_\star\right\|_{\mathrm{op}}$, combining these inequalities yields

$$D_5 \leqslant \left(\frac{\left\|Z^t\right\|_{\mathrm{op}}^2}{\left\|Z_\star\right\|_{\mathrm{op}}^2}\frac{\zeta_n^2 k}{2} + \frac{\zeta_n^2}{4} + \frac{\zeta_n^2 k}{2}\right) + \left(\frac{2\left\|Z^t\right\|_{\mathrm{op}}^2}{\left\|Z_\star\right\|_{\mathrm{op}}^2} + \frac{3}{2}\right)\left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2.$$

By the assumption that $\left\|\Delta_{Z^t}\right\|_{\mathrm{F}} \leqslant \frac{c_0}{e^{M_1}\kappa^2}\left\|Z_\star\right\|_{\mathrm{op}}$ for some sufficiently small c_0,

$$D_5 \leqslant C_5\left(\zeta_n^2 k + \left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2\right). \tag{44}$$

Combining (40), (41), (42), (43) and (44), we obtain

$$\begin{aligned}\widetilde{e}_{t+1} \leqslant \widetilde{e}_t &- \eta\left(\frac{2(\sqrt{2}-1)}{\kappa^2}\left(\frac{(1-c_1)\tau}{4\tau+1} - c_2\right) - \frac{1}{2C_0 e^{M_1}\kappa^2} + \frac{c_0}{4c_3 e^{M_1}\kappa^2}\right)e_t + \eta\left(C_1 + \frac{1}{4c_4}\right)\left\|\bar{G}_k\right\|_{\mathrm{F}}^2 \\ &- \left(\frac{1}{\tau+1/4} - c_3 - c_4 - C_5\eta\right)\left\|\sigma(\Theta^t) - \sigma(\Theta_\star)\right\|_{\mathrm{F}}^2 + \eta\frac{C_2}{4c_2}\zeta_n^2 k + \eta^2 C_5\zeta_n^2 k\end{aligned}$$

where c_2, c_3, c_4 are arbitrary constants, c_0 is a sufficiently small constant, and C_0 is a sufficiently large constant. Notice that $\tau \asymp e^{-M_1}$. Choose $c_2 = c\tau$ and c, c_3, c_4, η small enough such that

$$2(\sqrt{2}-1)\left(\frac{(1-c_1)\tau}{4\tau+1} - c_2\right) - \frac{1}{2e^{M_1}C_0} - \frac{c_0}{4c_3 e^{M_1}} > \widetilde{\rho}e^{-M_1}, \quad \text{and}$$
$$\frac{1}{\tau+1/4} - c_3 - c_4 - C_5\eta \geqslant 0,$$

for some positive constant $\widetilde{\rho}$. Recall that $\widetilde{e}_t \geqslant (1-\delta)e_t$. Then there exists a universal constant $C > 0$ such that

$$\widetilde{e}_{t+1} \leqslant \left(1 - \frac{\eta}{e^{M_1}\kappa^2}\widetilde{\rho}(1-\delta)\right)\widetilde{e}_t + \eta C\left(\left\|\bar{G}_k\right\|_{\mathrm{F}}^2 + e^{M_1}\zeta_n^2 k\right).$$

The proof is completed by setting $\rho = (1-\delta)\widetilde{\rho}$.

8.2.3 Proof of Lemma 8.7

Note the claim is deterministic in nature and we prove by induction. At initialization we have

$$\left\|\Delta_{Z^0}\right\|_{\mathrm{F}} \leqslant \left(\frac{\widetilde{e}_0}{\left\|Z^0\right\|_{\mathrm{op}}^2}\right)^{\frac{1}{2}} \leqslant \left(\frac{c_0^2}{4e^{2M_1}\kappa^4}\frac{\left\|Z_\star\right\|_{\mathrm{op}}^4}{\left\|Z^0\right\|_{\mathrm{op}}^2}\right)^{\frac{1}{2}} = \frac{c_0}{2e^{M_1}\kappa^2}\left\|Z_\star\right\|_{\mathrm{op}}\frac{\left\|Z_\star\right\|_{\mathrm{op}}}{\left\|Z^0\right\|_{\mathrm{op}}} \leqslant \frac{c_0}{e^{M_1}\kappa^2}\left\|Z_\star\right\|_{\mathrm{op}},$$

where the last inequality is obtained from

$$\left\|Z^0\right\|_{\mathrm{op}} \geqslant \left\|Z_\star\right\|_{\mathrm{op}} - \left\|\Delta_{Z^0}\right\|_{\mathrm{op}} \geqslant \left(1 - \frac{c_0}{2e^{M_1}\kappa^2}\right)\left\|Z_\star\right\|_{\mathrm{op}} \geqslant \frac{3}{4}\left\|Z_\star\right\|_{\mathrm{op}},$$

where the second the the last inequalities are due to Assumption 4.2.

Suppose the claim is true for all $t \leqslant t_0$, by Lemma 8.6,

$$\begin{aligned}
\widetilde{e}_{t_0+1} &\leqslant \left(1 - \frac{\eta}{e^{M_1}\kappa^2}\rho\right)^{t_0} \widetilde{e}_0 + \eta C \left(\|\bar{G}_k\|_{\mathrm{F}}^2 + e^{M_1}\zeta_n^2 k\right) \\
&\leqslant \widetilde{e}_0 + \eta C \left(\|\bar{G}_k\|_{\mathrm{F}}^2 + e^{M_1}\zeta_n^2 k\right) \\
&\leqslant \frac{c_0^2}{4e^{2M_1}\kappa^4} \|Z_\star\|_{\mathrm{op}}^4 + \eta C \left(\|\bar{G}_k\|_{\mathrm{F}}^2 + e^{M_1}\zeta_n^2 k\right) \\
&= \frac{c_0^2}{e^{2M_1}\kappa^4} \|Z_\star\|_{\mathrm{op}}^4 \left(\frac{1}{4} + \eta \frac{C e^{2M_1}\zeta_n^2 \kappa^4}{c_0^2 \|Z_\star\|_{\mathrm{op}}^4} \left(\frac{\|\bar{G}_k\|_{\mathrm{F}}^2}{\zeta_n^2} + e^{M_1} k\right)\right) \\
&\leqslant \frac{c_0^2}{e^{2M_1}\kappa^4} \|Z_\star\|_{\mathrm{op}}^4 \left(\frac{1}{4} + \frac{C}{c_0^2 C_1^2}\right).
\end{aligned}$$

Choosing C_1 large enough such that $C_1^2 \geqslant \frac{4C}{c_0^2}$, then

$$\widetilde{e}_{t_0+1} \leqslant \frac{c_0^2}{2e^{2M_1}\kappa^4} \|Z_\star\|_{\mathrm{op}}^4$$

and therefore,

$$\|\Delta_{Z^{t_0+1}}\|_{\mathrm{F}} \leqslant \left(\frac{\widetilde{e}_{t_0+1}}{\|Z_\star\|_{\mathrm{op}}^2}\right)^{\frac{1}{2}} \leqslant \frac{c_0}{\sqrt{2}e^{M_1}\kappa^2} \|Z_\star\|_{\mathrm{op}} \frac{\|Z_\star\|_{\mathrm{op}}}{\|Z^0\|_{\mathrm{op}}} \leqslant \frac{c_0}{e^{M_1}\kappa^2} \|Z_\star\|_{\mathrm{op}}.$$

This completes the proof.

8.3 Additional technique lemmas

We state below additional technical lemmas used in the proofs.

Lemma 8.8 ([21]). *Let $X_1, \cdots, X_n$ be independent Bernoulli random variables with $P(X_i = 1) = p_i$. For $S_n = \sum_{i=1}^n a_i X_i$ and $\nu = \sum_{i=1}^n a_i^2 p_i$. Then we have*

$$\begin{aligned}
P(S_n - \mathbb{E}S_n < -\lambda) &\leqslant \exp(-\lambda^2/2\nu), \\
P(S_n - \mathbb{E}S_n > \lambda) &\leqslant \exp\left(-\frac{\lambda^2}{2(\nu + a\lambda/3)}\right),
\end{aligned}$$

where $a = \max\{|a_1|, \cdots, |a_n|\}$.

Lemma 8.9 ([60]). *For any $Z_1, Z_2 \in \mathbb{R}^{n\times k}$, we have*

$$\mathrm{dist}(Z_1, Z_2)^2 \leqslant \frac{1}{2(\sqrt{2}-1)\sigma_k^2(Z_1)} \|Z_1 Z_1^\top - Z_2 Z_2^\top\|_{\mathrm{F}}^2.$$

Lemma 8.10 ([60]). *For any $Z_1, Z_2 \in \mathbb{R}^{n\times k}$ such that $\mathrm{dist}(Z_1, Z_2) \leqslant c\|Z_1\|_{\mathrm{op}}$, we have*

$$\|Z_1 Z_1^\top - Z_2 Z_2^\top\|_{\mathrm{F}} \leqslant (2 + c)\|Z_1\|_{\mathrm{op}} \mathrm{dist}(Z_1, Z_2).$$

Lemma 8.11 ([50]). *For a continuously differentiable function f, if it is μ-strongly convex and L-smooth on a convex domain $\mathcal{D}$, say for any $x, y \in \mathcal{D}$,*

$$\frac{\mu}{2}\|x-y\|^2 \leqslant f(y) - f(x) - \big\langle f'(x), y-x\big\rangle \leqslant \frac{L}{2}\|x-y\|^2,$$

then

$$\big\langle f'(x) - f'(y), x-y\big\rangle \geqslant \frac{\mu L}{\mu+L}\|x-y\|^2 + \frac{1}{\mu+L}\|f'(x)-f'(y)\|^2,$$

and also

$$\big\langle f'(x) - f'(y), x-y\big\rangle \geqslant \mu\|x-y\|^2.$$

Lemma 8.12 ([45], [26]). *Let A be the symmetric adjacency matrix of a random graph on n nodes in which edges occur independently. Let $\mathbb{E}[A_{ij}] = P_{ij}$ for all $i \neq j$ and $P_{ii} \in [0,1]$. Assume that $n \max_{i,j} P_{ij} \leqslant d$. Then for any C_0, there is a constant $C = C(C_0)$ such that*

$$\|A-P\|_{\mathrm{op}} \leqslant C\sqrt{d+\log n}$$

with probability at least $1 - n^{-C_0}$.

Lemma 8.13. *Let A be the symmetric adjacency matrix of a random graph of n nodes in which edges occur independently. Let $\mathbb{E}[A_{ij}] = P_{ij}$ for all $i \neq j$ and $P_{ii} \in [0,1]$ for all i and X be deterministic with $X_{ii} = 0$ for all i. Then,*

$$|\langle A-P, X\rangle| \leqslant C\|X\|_{\mathrm{F}}$$

with probability at least $1 - 2\exp(-C^2/8p_{\max}) - \exp(-C^2\|X\|_{\mathrm{F}}/8\|X\|_\infty)$, where $p_{\max} = \max_{i\neq j} P_{ij}$.

Proof. Observe that $\langle A-P, X\rangle = 2\sum_{i<j}(A_{ij}-P_{ij})X_{ij}$ and A_{ij} are independent Bernoulli random variables with $\mathbb{E}[A_{ij}] = P_{ij}$. Apply Lemma 8.8 to $\sum_{i<j}(A_{ij}-P_{ij})X_{ij}$ with $\lambda = C\|X\|_{\mathrm{F}}/2$, we have $\nu = \sum_{i<j} X_{ij}^2 P_{ij} \leqslant p_{\max}\|X\|_{\mathrm{F}}^2$ and

$$\begin{aligned} P\big(|\langle A-P, X\rangle| \leqslant C\|X\|_{\mathrm{F}}\big) &\leqslant \exp(-C^2\|X\|_{\mathrm{F}}^2/8\nu) + \exp\left(-\frac{C^2\|X\|_{\mathrm{F}}^2}{8\max\{\nu, C\|X\|_\infty\|X\|_{\mathrm{F}}\}}\right) \\ &\leqslant 2\exp(-C^2\|X\|_{\mathrm{F}}^2/8\nu) + \exp(-C^2\|X\|_{\mathrm{F}}/8\|X\|_\infty) \\ &\leqslant 2\exp(-C^2/8p_{\max}) + \exp(-C^2\|X\|_{\mathrm{F}}/8\|X\|_\infty). \end{aligned}$$

This completes the proof. □

A Proofs of results for initialization

This section presents the proofs of Theorem 4.5, Corollary 4.1 and Proposition 4.1.

A.1 Preliminaries

We introduce two technical results used repeatedly in the proofs.

Lemma A.1. *If $\|G_\star\|_{\mathrm{op}}^2 \geqslant C\|\bar{G}_k\|_{\mathrm{F}}^2$ for some constant $C > 0$, then $\|G_\star\|_{\mathrm{op}}^2 \geqslant c\|G_\star\|_{\mathrm{F}}^2/k$ for some constant $c > 0$.*

Proof. By definition,

$$\|G_\star\|_{\mathrm{F}}^2 \leqslant 2\left(\|\bar{G}_k\|_{\mathrm{F}}^2 + \|Z_\star Z_\star^\top\|_{\mathrm{F}}^2\right) \leqslant 2\|\bar{G}_k\|_{\mathrm{F}}^2 + 2k\,\|Z_\star\|_{\mathrm{op}}^4 \leqslant (2k + 1/C)\,\|Z_\star\|_{\mathrm{op}}^4 .$$

Therefore, $\|Z_\star\|_{\mathrm{op}}^4 \geqslant c\|G_\star\|_{\mathrm{F}}^2/k$ for some constant $c > 0$. □

Theorem A.1. *Under Assumption 4.1, choose λ_n, γ_n such that*

$$\lambda_n \geqslant 2\max\left\{\|A-P\|_{\mathrm{op}} + \gamma_n\|G_\star\|_{\mathrm{op}}, \|A-P\|_{\mathrm{op}} + \frac{\gamma_n}{\sqrt{k}}\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} + \frac{\gamma_n}{\sqrt{k}}\|X\beta_\star\|_{\mathrm{F}}\right\}. \tag{45}$$

Let the constant step size $\eta \leqslant 2/9$ and the constraint sets $\mathcal{C}_G, \mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ as specified in Theorem 4.5. If the latent vectors contain strong enough signal in the sense that

$$\|G_\star\|_{\mathrm{op}}^2 \geqslant C\kappa_{Z_\star}^6 e^{2M_1}\max\left\{e^{2M_1}\lambda_n^2 k,\ \|\bar{G}_k\|_*^2/k,\ \|\bar{G}_k\|_{\mathrm{F}}^2\right\} \tag{46}$$

for some sufficiently large constant C, then for any given constant $c_1 > 0$, there exists a universal constant C_1 such that for any $T \geqslant T_0$, the error will satisfy $e_T \leqslant c_1^2 e^{-2M_1}\|Z_\star\|_{\mathrm{op}}^4/\kappa_{Z_\star}^4$, where

$$T_0 = \log\left(\frac{C_1 e^{2M_1} k\kappa_{Z_\star}^6}{c_1^2}\frac{\|G_\star\|_{\mathrm{F}}^2 + 2\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}^2 + \|X\beta_\star\|_{\mathrm{F}}^2}{\|G_\star\|_{\mathrm{F}}^2}\right)\left(\log\left(\frac{1}{1-\gamma_n\eta}\right)\right)^{-1}.$$

Proof. See Section A.5. □

A.2 Proof of Theorem 4.5

We focus on the case where X is nonzero, and the case of $X = 0$ is simpler. By Lemma 8.3, there exist constants C_2, c such that with probability at least $1 - n^c$,

$$\|A-P\|_{\mathrm{op}},\ \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} \leqslant C_2\sqrt{\max\{ne^{-M_2}, \log n\}}.$$

All the following analysis is conditional on this event. Since $\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \|\beta_\star X\|_{\mathrm{F}} \leqslant C\|G_\star\|_{\mathrm{F}}$, by Lemma A.1,

$$\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}} + \|X\beta_\star\|_{\mathrm{F}} \leqslant C_3\sqrt{k}\,\|G_\star\|_{\mathrm{op}} .$$

for some constant C_3. Combining these two inequalities leads to

$$\begin{aligned}&\max\left\{\|A-P\|_{\mathrm{op}} + \gamma_n\|G_\star\|_{\mathrm{op}}, \|A-P\|_{\mathrm{op}} + \frac{\gamma_n}{\sqrt{k}}\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} + \frac{\gamma_n}{\sqrt{k}}\|X\beta_\star\|_{\mathrm{F}}\right\}\\ &\leqslant C_2\sqrt{\max\{ne^{-M_2}, \log n\}} + (1+C_3)\gamma_n\|G_\star\|_{\mathrm{op}} \leqslant C_2/C_0\lambda_n + (1+C_3)\delta\lambda_n \leqslant \lambda_n/2.\end{aligned}$$

Here the last inequality is due to fact that C_0 is sufficiently large and δ is sufficiently small. Furthermore,

$$\widetilde{C}\kappa_{Z_\star}^6 e^{4M_1}\lambda_n^2 k \leqslant \widetilde{C}c_0^2 \left\|G_\star\right\|_{\mathrm{op}}^2 \leqslant \left\|G_\star\right\|_{\mathrm{op}}^2,$$

since c_0 is a sufficient small constant. Therefore, the inequality (46) holds. Apply Theorem A.1, there exists a universal constant C_1 such that for any given constant $c_1 > 0$, $e_T \leqslant c_1^2 e^{-2M_1}\left\|Z_\star\right\|_{\mathrm{op}}^4/\kappa_{Z_\star}^4$, as long as $T \geqslant T_0$, where

$$T_0 = \log\left(\frac{C_1 e^{2M_1} k\kappa_{Z_\star}^6}{c_1^2}\frac{\|x_\star\|_{\mathcal{D}}^2}{\|G_\star\|_{\mathrm{F}}^2}\right)\left(\log\left(\frac{1}{1-\gamma_n\eta}\right)\right)^{-1}.$$

Notice that when $\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \|\beta_\star X\|_{\mathrm{F}} \leqslant C\|G_\star\|_{\mathrm{F}}$, $\|x_\star\|_{\mathcal{D}}^2 \leqslant C_4\|G_\star\|_{\mathrm{F}}^2$ for some constant C_4. Therefore,

$$T_0 \leqslant \log\left(\frac{C_1 C_4 e^{2M_1} k\kappa_{Z_\star}^6}{c_1^2}\right)\left(\log\left(\frac{1}{1-\gamma_n\eta}\right)\right)^{-1}.$$

This completes the proof.

A.3 Proof of Corollary 4.1

By Lemma 8.3, there exist constants C_2, c_2 such that with probability at least $1 - n^{c_2}$,

$$\|A-P\|_{\mathrm{op}},\ \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} \leqslant C_2\sqrt{\max\{ne^{-M_2}, \log n\}}.$$

All the following analysis is conditional on this event. Since $\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \|\beta_\star X\|_{\mathrm{F}} \leqslant C\|G_\star\|_{\mathrm{F}}$, by Lemma A.1,

$$\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}} + \|X\beta_\star\|_{\mathrm{F}} \leqslant C_3\sqrt{k}\left\|G_\star\right\|_{\mathrm{op}}.$$

for some constant C_3. Combining these two inequalities leads to

$$\begin{aligned}
&\max\left\{\|A-P\|_{\mathrm{op}} + \gamma_n\left\|G_\star\right\|_{\mathrm{op}}, \|A-P\|_{\mathrm{op}} + \frac{\gamma_n}{\sqrt{k}}\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} + \frac{\gamma_n}{\sqrt{k}}\|X\beta_\star\|_{\mathrm{F}}\right\}\\
&\leqslant C_2\sqrt{\max\{ne^{-M_2}, \log n\}} + (1+C_3)\gamma_n\left\|G_\star\right\|_{\mathrm{op}}\\
&\leqslant C_2\gamma_n\left\|G_\star\right\|_{\mathrm{op}} + (1+C_3)\gamma_n\left\|G_\star\right\|_{\mathrm{op}} = (1+C_2+C_3)\gamma_n\left\|G_\star\right\|_{\mathrm{op}},
\end{aligned}$$

where the last inequality is due to equation (22). Since we choose $\lambda_n = C_0\gamma_n\left\|Z_\star\right\|_{\mathrm{op}}^2$ for some sufficiently large constant C_0, inequality (45) holds. Further, notice that $\gamma_n = \gamma = c_0/(e^{2M_1}\sqrt{k}\kappa_{Z_\star}^3)$ for some sufficiently small constant c_0,

$$\widetilde{C}e^{4M_1}\kappa_{Z_\star}^6\lambda_n^2 k = \widetilde{C}c_0^2\left\|G_\star\right\|_{\mathrm{op}}^2 \leqslant \left\|G_\star\right\|_{\mathrm{op}}^2.$$

Therefore, the inequality (46) holds. Apply Theorem A.1, there exists a universal constant C_1 such that for any given constant $c_1 > 0$, $e_T \leqslant c_1^2 e^{-2M_1}\left\|Z_\star\right\|_{\mathrm{op}}^4/\kappa_{Z_\star}^4$, as long as $T \geqslant T_0$, where

$$T_0 = \log\left(\frac{C_1 e^{2M_1} k\kappa_{Z_\star}^6}{c_1^2}\frac{\|x_\star\|_{\mathcal{D}}^2}{\|G_\star\|_{\mathrm{F}}^2}\right)\left(\log\left(\frac{1}{1-\gamma\eta}\right)\right)^{-1}.$$

Notice that when $\|\alpha_\star 1_n^\top\|_{\mathrm{F}}, \|\beta_\star X\|_{\mathrm{F}} \leqslant C\|G_\star\|_{\mathrm{F}}$, $\|x_\star\|_{\mathcal{D}}^2 \leqslant C_4\|G_\star\|_{\mathrm{F}}^2$ for some constant C_4. Therefore,

$$T_0 \leqslant \log\left(\frac{C_1C_4e^{2M_1}k\kappa_{Z_\star}^6}{c_1^2}\right)\left(\log\left(\frac{1}{1-\gamma\eta}\right)\right)^{-1}.$$

This completes the proof.

A.4 Proof of Proposition 4.1

Applying Theorem 2.7 in [16] we obtain

$$\frac{1}{n^2}\|\widehat{P}-P\|_{\mathrm{F}}^2 \leqslant C(k,M_1,\kappa_{Z_\star})n^{-\frac{1}{k+3}}.$$

where the constant $C(k,M_1,\kappa_{Z_\star})$ depends on $k, M_1, \kappa_{Z_\star}$. Notice that $\Theta_{ij} = \mathrm{logit}(P_{ij})$ and $\mathrm{logit}(\cdot)$ is $4e^{M_1}$-Lipchitz continuous in the interval $\left[\frac{1}{2}e^{-M_1}, \frac{1}{2}\right]$, and so

$$\frac{1}{n^2}\|\widehat{\Theta}-\Theta\|_{\mathrm{F}}^2 \leqslant C'(k,M_1,\kappa_{Z_\star})n^{-\frac{1}{k+3}}.$$

Let $\Delta_{\widehat{\Theta}} = \widehat{\Theta} - \Theta_\star$ It is easy to verify,

$$\alpha^0 = \left(2nI_n + 21_n1_n^\top\right)^{-1}\widehat{\Theta}1_n = \alpha_\star + \frac{1}{n}\left(I_n - \frac{1}{2n}1_n1_n^\top\right)\Delta_{\widehat{\Theta}}1_n,$$

and hence

$$\begin{aligned}\|\alpha^0 1_n^\top - \alpha_\star 1_n^\top\|_{\mathrm{F}} &= \frac{1}{n}\left\|\left(I_n - \frac{1}{2n}1_n1_n^\top\right)\Delta_{\widehat{\Theta}}1_n1_n^\top\right\|_{\mathrm{F}} \\ &\leqslant \frac{1}{n}\left\|I_n - \frac{1}{2n}1_n1_n^\top\right\|_{\mathrm{op}}\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}\|1_n1_n^\top\|_{\mathrm{F}} \leqslant \|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}.\end{aligned}$$

Notice that $G_\star \in \mathbb{S}_+^n$,

$$\|\widehat{G}-G_\star\|_{\mathrm{F}} \leqslant \|\widehat{G}-J\widehat{\Theta}J + J\widehat{\Theta}J - G_\star\|_{\mathrm{F}} \leqslant 2\|J\widehat{\Theta}J - G_\star\|_{\mathrm{F}} \leqslant 2\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}.$$

Further notice that $\mathrm{r}(G_\star) = k$,

$$\|Z^0(Z^0)^\top - G_\star\|_{\mathrm{F}} \leqslant \|Z^0(Z^0)^\top - \widehat{G} + \widehat{G} - G_\star\|_{\mathrm{F}} \leqslant 2\|\widehat{G}-G_\star\|_{\mathrm{F}} \leqslant 4\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}.$$

Then, by Lemma 8.9,

$$\begin{aligned}e_0 &\leqslant \|Z_\star\|_{\mathrm{op}}^2\,\mathrm{dist}(Z^0,Z_\star)^2 + 2n\left\|\alpha^0-\alpha_\star\right\|^2 \\ &\leqslant \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)}\|Z^0(Z^0)^\top - G_\star\|_{\mathrm{F}}^2 + 2n\left\|\alpha^0-\alpha_\star\right\|^2 \leqslant 24\kappa_{Z_\star}^2\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widehat{\Theta}}\|_{\mathrm{F}}^2 \\ &\leqslant 26\kappa_{Z_\star}^2C'(k,M_1,\kappa_{Z_\star})n^2 \times n^{-\frac{1}{k+3}} \leqslant \frac{26k\kappa_{Z_\star}^2C'(k,M_1,\kappa_{Z_\star})}{c_0}\frac{\|G_\star\|_{\mathrm{F}}^2}{k} \times n^{-\frac{1}{k+3}} \\ &\leqslant C_1(k,M_1,\kappa_{Z_\star})\,\|Z_\star\|_{\mathrm{op}}^4 \times n^{-\frac{1}{k+3}}.\end{aligned}$$

Therefore, the initialization condition in Assumption 4.2 will hold for large enough n.

A.5 Proof of Theorem A.1

A.5.1 Preparations

Recall the definition of $f(G, \alpha, \beta)$ in (15) where

$$(G, \alpha, \beta) \in \mathcal{D} = \left\{ (G, \alpha, \beta) | GJ = G, G \in \mathbb{S}_+, \max_{i,j} |G_{ij}|, \max_i |\alpha_i| \leqslant \frac{M}{3}, |\beta| \leqslant \frac{M}{3 \max_{i,j} |X_{ij}|} \right\}.$$

Define the norm $\|\cdot\|_{\mathcal{D}}$ in the domain $\mathcal{D}$ by

$$\|(G, \alpha, \beta)\|_{\mathcal{D}} = \left(\|G\|_{\mathrm{F}}^2 + 2\|\alpha 1_n{}^\top\|_{\mathrm{F}}^2 + \|X\beta\|_{\mathrm{F}}^2 \right)^{1/2}.$$

Lemma A.2. *The function f is γ_n-strongly convex and $(\gamma_n + 9/2)$-smooth in the convex domain $\mathcal{D}$ with respect to the norm $\|\cdot\|_{\mathcal{D}}$, that is, for $(G_i, \alpha_i, \beta_i) \in \mathcal{D}, i = 1, 2$, let $(\Delta_G, \Delta_\alpha, \Delta_\beta) = (G_1 - G_2, \alpha_1 - \alpha_2, \beta_1 - \beta_2)$, then*

$$\begin{aligned} \frac{\gamma_n}{2} \|(\Delta_G, \Delta_\alpha, \Delta_\beta)\|_{\mathcal{D}}^2 \leqslant\ & f(G_1, \alpha_1, \beta_1) - f(G_2, \alpha_2, \beta_2) - \left\langle \nabla_G f(G_2, \alpha_2, \beta_2), \Delta_G \right\rangle \\ & - \left\langle \nabla_\alpha f(G_2, \alpha_2, \beta_2), \Delta_\alpha \right\rangle - \left\langle \nabla_\beta f(G_2, \alpha_2, \beta_2), \Delta_\beta \right\rangle \\ \leqslant\ & \frac{\gamma_n + 9/2}{2} \|(\Delta_G, \Delta_\alpha, \Delta_\beta)\|_{\mathcal{D}}^2. \end{aligned}$$

Proof. With slight abuse of notation, let

$$h(G, \alpha, \beta) = -\sum_{i,j} \left\{ A_{ij}\Theta_{ij} + \log\left(1 - \sigma(\Theta_{ij})\right) \right\} \tag{47}$$

which is a convex function of G, α and β. In addition, let

$$r(G, \alpha, \beta) = \frac{\gamma_n}{2} \left(\|G\|_{\mathrm{F}}^2 + 2\|\alpha 1_n{}^\top\|_{\mathrm{F}}^2 + \|X\beta\|_{\mathrm{F}}^2 \right) + \lambda_n \operatorname{Tr}(G) \tag{48}$$

which is γ_n-strongly convex w.r.t. the norm $\|\cdot\|_{\mathcal{D}}$. Thus $f(G, \alpha, \beta)$ is γ_n-strongly convex. On the other hand, $r(\cdot, \cdot, \cdot)$ is γ_n smooth and

$$\begin{aligned} & h(G_1, \alpha_1, \beta_1) - h(G_2, \alpha_2, \beta_2) - \left\langle \nabla_G h(G_2, \alpha_2, \beta_2), \Delta_G \right\rangle \\ & \qquad - \left\langle \nabla_\alpha h(G_2, \alpha_2, \beta_2), \Delta_\alpha \right\rangle - \left\langle \nabla_\beta h(G_2, \alpha_2, \beta_2), \Delta_\beta \right\rangle \\ = {} & h(\Theta_1) - h(\Theta_2) - \left\langle \nabla_\Theta h(\Theta_2), \Delta_G \right\rangle - \left\langle 2\nabla_\Theta h(\Theta_2) 1_n, \Delta_\alpha \right\rangle - \left\langle \nabla_\Theta h(\Theta_2), X \right\rangle \Delta_\beta \\ = {} & \frac{1}{2} h(\Theta_1) - h(\Theta_2) - \left\langle \nabla_\Theta h(\Theta_2), \Theta_1 - \Theta_2 \right\rangle \\ \leqslant {} & \frac{1}{8} \|\Theta_1 - \Theta_2\|_{\mathrm{F}}^2 = \frac{1}{8} \|\Delta_G + 2\Delta_\alpha 1_n{}^\top + X\Delta_\beta\|_{\mathrm{F}}^2 \\ \leqslant {} & \frac{9}{8} \left(\|\Delta_G\|_{\mathrm{F}}^2 + 4\|\Delta_\alpha 1_n{}^\top\|_{\mathrm{F}}^2 + \|X\Delta_\beta\|_{\mathrm{F}}^2 \right) \leqslant \frac{9}{4} \left(\|\Delta_G\|_{\mathrm{F}}^2 + 2\|\Delta_\alpha 1_n{}^\top\|_{\mathrm{F}}^2 + \|X\Delta_\beta\|_{\mathrm{F}}^2 \right). \end{aligned}$$

This finishes the proof. □

Define $(\widetilde{G},\ \widetilde{\alpha},\ \widetilde{\beta}) = \arg\min_{(G,\alpha,\beta)\in\mathcal{D}}\ \ f(G,\alpha,\beta)$, $\Delta_{\widetilde{G}} = \widetilde{G} - G_\star, \Delta_{\widetilde{\alpha}} = \widetilde{\alpha} - \alpha_\star, \Delta_{\widetilde{\beta}} = \widetilde{\beta} - \beta_\star, \Delta_{\widetilde{\Theta}} = \widetilde{\Theta} - \Theta_\star$. Similar to the analysis of the convex programming in (11), one can obtain the following results.

Lemma A.3. *Let $\mathcal{M}_k^\perp = \{M \in \mathbb{R}^{n\times n} : \mathrm{row}(M) \subset \mathrm{col}(Z_\star)^\perp \text{ and } \mathrm{col}(M) \subset \mathrm{col}(Z_\star)^\perp\}$ and $\mathcal{M}_k$ be its orthogonal complement in $\mathbb{R}^{n\times n}$ under trace inner product. If*

$$\lambda_n \geqslant 2\max\left\{\|A-P\|_{\mathrm{op}} + \gamma_n\|G_\star\|_{\mathrm{op}}, \|A-P\|_{\mathrm{op}} + \frac{\gamma_n}{\sqrt{k}}\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} + \frac{\gamma_n}{\sqrt{k}}\|X\beta_\star\|_{\mathrm{F}}\right\},$$

then for $\bar{G}_k = \mathcal{P}_{\mathcal{M}_k^\perp} G_\star$, we have

$$\|\Delta_{\widetilde{G}}\|_* \leqslant 4\sqrt{2k}\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widetilde{G}}\|_{\mathrm{F}} + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}}1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}} + 4\|\bar{G}_k\|_*.$$

Proof. Let

$$\widetilde{h}(G,\alpha,\beta) = -\sum_{1\leqslant i,j\leqslant n}\{A_{ij}\Theta_{ij} + \log(1-\sigma(\Theta_{ij}))\} + \frac{\gamma_n}{2}\left(\|G\|_{\mathrm{F}}^2 + 2\|\alpha 1_n{}^\top\|_{\mathrm{F}}^2 + \|X\beta\|_{\mathrm{F}}^2\right).$$

By the convexity of $\widetilde{h}$,

$$\begin{aligned}
&\widetilde{h}(\widetilde{G},\widetilde{\alpha},\widetilde{\beta}) - \widetilde{h}(G_\star,\alpha_\star,\beta_\star)\\
&\geqslant \left\langle \nabla_G\widetilde{h}(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{G}}\right\rangle + \left\langle \nabla_\alpha\widetilde{h}(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{\alpha}}\right\rangle + \left\langle \nabla_\beta\widetilde{h}(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{\beta}}\right\rangle\\
&= -\langle A-P,\ \Delta_{\widetilde{G}} + 2\Delta_{\widetilde{\alpha}}1_n{}^\top + \Delta_{\widetilde{\beta}}X\rangle + \gamma_n\left(\left\langle G_\star, \Delta_{\widetilde{G}}\right\rangle + 2n\left\langle \alpha_\star, \Delta_{\widetilde{\alpha}}\right\rangle + \|X\|_{\mathrm{F}}^2\left\langle \beta_\star, \Delta_{\widetilde{\beta}}\right\rangle\right)\\
&\geqslant -\|A-P\|_{\mathrm{op}}\left(\|\Delta_{\widetilde{G}}\|_* + 2\|\Delta_{\widetilde{\alpha}}1_n{}^\top\|_*\right) - |\langle A-P, \Delta_{\widetilde{\beta}}X\rangle|\\
&\quad - \gamma_n\left(\|G_\star\|_{\mathrm{op}}\|G_\star\|_* + 2\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}\|\Delta_{\widetilde{\alpha}}1_n{}^\top\|_{\mathrm{F}} + \|X\beta_\star\|_{\mathrm{F}}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\right)\\
&\geqslant -\left(\|A-P\|_{\mathrm{op}} + \gamma_n\|G_\star\|_{\mathrm{op}}\right)\|\Delta_{\widetilde{G}}\|_* - \left(\|A-P\|_{\mathrm{op}} + \gamma_n/\sqrt{k}\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}\right)2\sqrt{k}\|\Delta_{\widetilde{\alpha}}1_n{}^\top\|_{\mathrm{F}}\\
&\quad - \left(\left\langle A-P, X\right\rangle/\left(\sqrt{k}\|X\|_{\mathrm{F}}\right) + \gamma_n/\sqrt{k}\|X\beta_\star\|_{\mathrm{F}}\right)\sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\\
&\geqslant -\frac{\lambda_n}{2}\left(\|\Delta_{\widetilde{G}}\|_* + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}}1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\right)\\
&\geqslant -\frac{\lambda_n}{2}\left(\|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widetilde{G}}\|_* + \|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widetilde{G}}\|_* + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}}1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\right).
\end{aligned}$$

The last inequality holds since $\mathcal{P}_{\mathcal{M}_k} + \mathcal{P}_{\mathcal{M}_k^\perp}$ equals identity and

$$\lambda_n \geqslant 2\max\left\{\|A-P\|_{\mathrm{op}} + \gamma_n\|G_\star\|_{\mathrm{op}}, \|A-P\|_{\mathrm{op}} + \frac{\gamma_n}{\sqrt{k}}\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \frac{\langle A-P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} + \frac{\gamma_n}{\sqrt{k}}\|X\beta_\star\|_{\mathrm{F}}\right\}.$$

On the other hand, by the definition of $\bar{G}_k$,

$$\begin{aligned}
\|\widetilde{G}\|_* - \|G_\star\|_* &= \|\mathcal{P}_{\mathcal{M}_k}G_\star + \bar{G}_k + \mathcal{P}_{\mathcal{M}_k}\Delta_{\widetilde{G}} + \mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widetilde{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k}G_\star + \bar{G}_k\|_*\\
&\geqslant \|\mathcal{P}_{\mathcal{M}_k}G_\star + \mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widetilde{G}}\|_* - \|\bar{G}_k\|_* - \|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widetilde{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k}G_\star\|_* - \|\bar{G}_k\|_*\\
&= \|\mathcal{P}_{\mathcal{M}_k}G_\star\|_* + \|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widetilde{G}}\|_* - 2\|\bar{G}_k\|_* - \|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widetilde{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k}G_\star\|_*\\
&= \|\mathcal{P}_{\mathcal{M}_k^\perp}\Delta_{\widetilde{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k}\Delta_{\widetilde{G}}\|_* - 2\|\bar{G}_k\|_*.
\end{aligned}$$

Here, the second last equality holds since $\mathcal{P}_{\mathcal{M}_k} G_\star$ and $\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widetilde{G}}$ have orthogonal column and row spaces. Furthermore, since $\widehat{\Theta}$ is the optimal solution to (11), and $\Theta_\star$ is feasible, the basic inequality and the last two displays imply

$$\begin{aligned}
0 &\geqslant \widetilde{h}(\widetilde{G}, \widetilde{\alpha}, \widetilde{\beta}) - \widetilde{h}(G_\star, \alpha_\star, \beta_\star) + \lambda_n\big(\|\widetilde{G}\|_* - \|G_\star\|_*\big) \\
&\geqslant -\frac{\lambda_n}{2}\big(\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widetilde{G}}\|_* + \|\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widetilde{G}}\|_* + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\big) \\
&\quad + \lambda_n\big(\|\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widetilde{G}}\|_* - \|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widetilde{G}}\|_* - 2\|\bar{G}_k\|_*\big) \\
&= \frac{\lambda_n}{2}\big(\|\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widetilde{G}}\|_* - 3\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widetilde{G}}\|_* - 4\|\bar{G}_k\|_* - 2\sqrt{k}\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} - \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\big).
\end{aligned}$$

Rearranging the terms leads to

$$\|\mathcal{P}_{\mathcal{M}_k^\perp} \Delta_{\widetilde{G}}\|_* \leqslant 3\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widetilde{G}}\|_* + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}} + 4\|\bar{G}_k\|_*\,,$$

and triangle inequality further implies

$$\|\Delta_{\widetilde{G}}\|_* \leqslant 4\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widetilde{G}}\|_* + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}} + 4\|\bar{G}_k\|_*\,.$$

Finally, note that the rank of $\mathcal{P}_{\mathcal{M}_k} \Delta_{\widetilde{G}}$ is at most $2k$,

$$\|\Delta_{\widetilde{G}}\|_* \leqslant 4\sqrt{2k}\|\mathcal{P}_{\mathcal{M}_k} \Delta_{\widetilde{G}}\|_{\mathrm{F}} + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}} + 4\|\bar{G}_k\|_*\,.$$

This completes the proof. □

Lemma A.4. *For any $k \geqslant 1$ such that Assumption 4.1 holds. Choose $\lambda_n \geqslant \max\{2\|A - P\|_{\mathrm{op}}, 1\}$ and $|\langle A - P, X\rangle| \leqslant \lambda_n\sqrt{k}\|X\|_{\mathrm{F}}$. There exist constants $C > 0$ and $0 \leqslant c < 1$ such that*

$$\begin{aligned}
\|\Delta_{\widetilde{\Theta}}\|_{\mathrm{F}}^2 &\geqslant (1-c)\big(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}^2\big) - C\|\bar{G}_k\|_*^2/k, \quad \text{and} \\
\|\Delta_{\widetilde{\Theta}}\|_{\mathrm{F}}^2 &\leqslant (1+c)\big(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}^2\big) + C\|\bar{G}_k\|_*^2/k.
\end{aligned}$$

Proof. The proof is the same as the proof of Lemma 8.2 and we leave out the details. □

Theorem A.2. *Under Assumption 4.1, for any λ_n satisfying*

$$\lambda_n \geqslant 2\max\left\{\|A - P\|_{\mathrm{op}} + \gamma_n\|G_\star\|_{\mathrm{op}}\,, \|A - P\|_{\mathrm{op}} + \frac{\gamma_n}{\sqrt{k}}\|\alpha_\star 1_n{}^\top\|_{\mathrm{F}}, \frac{\langle A - P, X\rangle}{\sqrt{k}\|X\|_{\mathrm{F}}} + \frac{\gamma_n}{\sqrt{k}}\|X\beta_\star\|_{\mathrm{F}}\right\},$$

there exists a constant C such that

$$\Big(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}\Big)^2 \leqslant C\left(e^{2M_1}\lambda_n^2 k + \frac{\|\bar{G}_k\|_*^2}{k}\right).$$

Proof. Recall the definition of $h(G, \alpha, \beta)$ in (47). Observe that $\widehat{\Theta} = \widehat{\alpha} 1_n{}^\top + 1_n\widehat{\alpha}^\top + \widehat{\beta}X + \widehat{G}$ is the optimal solution to (11), and that the true parameter $\Theta_\star = \alpha_\star 1_n{}^\top + 1_n\alpha_\star^\top + \beta_\star X + G_\star$ is feasible. Thus, we have the basic inequality

$$\widetilde{h}(\widetilde{G}, \widetilde{\alpha}, \widetilde{\beta}) - \widetilde{h}(G_\star, \alpha_\star, \beta_\star) + \lambda_n(\|\widetilde{G}\|_* - \|G_\star\|_*) \leqslant 0. \tag{49}$$

By definition,

$$\widetilde{h}(G,\alpha,\beta) = h(G,\alpha,\beta) + \frac{\gamma_n}{2}\left(\|G\|_{\mathrm{F}}^2 + \|\alpha 1_n{}^\top\|_2^2 + \|X\beta\|_{\mathrm{F}}^2\right).$$

On the one hand,

$$\begin{aligned}
&h(\widetilde{G},\widetilde{\alpha},\widetilde{\beta}) - h(G_\star,\alpha_\star,\beta_\star) \\
&\quad - \left\langle \nabla_G h(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{G}} \right\rangle - \left\langle \nabla_\alpha h(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{\alpha}} \right\rangle - \left\langle \nabla_\beta h(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{\beta}} \right\rangle \\
&= h(\widetilde{\Theta}) - h(\Theta_\star) - \left\langle \nabla_\Theta h(\Theta_\star), \Delta_{\widetilde{\Theta}} \right\rangle \geqslant \frac{\tau}{2}\|\Delta_{\widetilde{\Theta}}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last inequality is by the strong convexity of $h(\cdot)$ with respect to Θ in the domain $\mathcal{F}_g$ and $\tau = e^{M_1}/(1+e^{M_1})^2$ as in the proof of Theorem 4.1. Further by Lemma A.4,

$$\frac{\tau}{2}\|\Delta_{\widetilde{\Theta}}\|_{\mathrm{F}}^2 \geqslant \frac{\tau(1-c)}{2}\big(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}^2\big) - \frac{C\tau}{2}\|\bar{G}_k\|_*^2/k.$$

On the other hand, the l_2 regularization term is strongly convex with respect to (G,α,β). Then we have

$$\begin{aligned}
&\widetilde{h}(\widetilde{G},\widetilde{\alpha},\widetilde{\beta}) - \widetilde{h}(G_\star,\alpha_\star,\beta_\star) \\
&\geqslant \left\langle \nabla_G \widetilde{h}(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{G}} \right\rangle + \left\langle \nabla_\alpha \widetilde{h}(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{\alpha}} \right\rangle + \left\langle \nabla_\beta \widetilde{h}(G_\star,\alpha_\star,\beta_\star), \Delta_{\widetilde{\beta}} \right\rangle \\
&\quad + \frac{\tau(1-c)}{2}\big(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}^2\big) - \frac{C\tau}{2}\|\bar{G}_k\|_*^2/k \\
&\geqslant -\frac{\lambda_n}{2}\big(\|\Delta_{\widetilde{G}}\|_* + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\big) \\
&\quad + \frac{\tau(1-c)}{2}\big(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}^2\big) - \frac{C\tau}{2}\|\bar{G}_k\|_*^2/k.
\end{aligned}$$

By triangle inequality,

$$\lambda_n(\|\widehat{G}\|_* - \|G_\star\|_*) \geqslant -\lambda_n\|\Delta_G\|_*.$$

Together with (49), the last two inequalities imply

$$\begin{aligned}
&\frac{\tau(1-c)}{2}\left(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}^2\right) \\
&\quad \leqslant \frac{\lambda_n}{2}\left(\|\Delta_{\widetilde{G}}\|_* + 2\sqrt{k}\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \sqrt{k}\|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\right) + \lambda_n\|\Delta_{\widetilde{G}}\|_* + \frac{C\tau}{2}\|\bar{G}_k\|_*^2/k.
\end{aligned}$$

By Lemma A.3,

$$\begin{aligned}
&\frac{\tau(1-c)}{2}\left(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}}^2 + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}^2\right) \\
&\quad \leqslant C_0\lambda_n\sqrt{k}\left(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\right) + C_1\lambda_n\|\bar{G}_k\|_* + \frac{C\tau}{2}\|\bar{G}_k\|_*^2/k.
\end{aligned}$$

This implies that there exists some constant c_0 such that

$$\begin{aligned}
&c_0\tau\left(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}\right)^2 \\
&\quad \leqslant C_0\lambda_n\sqrt{k}\left(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \|X\Delta_{\widetilde{\beta}}\|_{\mathrm{F}}\right) + C_1\lambda_n\|\bar{G}_k\|_* + \frac{C\tau}{2}\|\bar{G}_k\|_*^2/k.
\end{aligned}$$

Solving the quadratic inequality, there exists some constant C_2 such that

$$\left(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}\right)^2 \leqslant C_2 \left(\frac{\lambda_n^2 k}{\tau^2} + \frac{\lambda_n \|\bar{G}_k\|_*}{\tau} + \frac{\|\bar{G}_k\|_*^2}{k}\right).$$

Note that $\tau \geqslant c_1 e^{-M_1}$ and $e^{M_1}\lambda_n\|\bar{G}_k\|_* \leqslant c_2\left(e^{2M_1}\lambda_n^2 k + \frac{\|\bar{G}_k\|_*^2}{k}\right)$ for positive constants c_1, c_2. Therefore,

$$\left(\|\Delta_{\widetilde{G}}\|_{\mathrm{F}} + 2\|\Delta_{\widetilde{\alpha}} 1_n{}^\top\|_{\mathrm{F}} + \|\Delta_{\widetilde{\beta}} X\|_{\mathrm{F}}\right)^2 \leqslant C_2 \left(e^{2M_1}\lambda_n^2 k + \frac{\|\bar{G}_k\|_*^2}{k}\right),$$

which completes the proof. □

Lemma A.5 ([10]). *Let $x \in \mathcal{D}$ and $y \in \mathbb{R}^n$, then*

$$\left\langle \pi_{\mathcal{D}}(y) - x, \pi_{\mathcal{D}}(y) - y \right\rangle \leqslant 0$$

where $\mathcal{D}$ is a convex set and $\pi_{\mathcal{D}}(x) = \arg\min_{y\in\mathcal{D}} \|x - y\|$.

Lemma A.6. *With $\eta_G = \eta, \eta_\alpha = \eta/2n, \eta_\beta = \eta/\|X\|_{\mathrm{F}}^2$,*

$$\begin{aligned}
&\left\langle G^t - G^{t+1}, G^t - \widetilde{G}\right\rangle + 2n\left\langle \alpha^t - \alpha^{t+1}, \alpha^t - \widetilde{\alpha}\right\rangle + \left\langle \beta^t - \beta^{t+1}, \beta^t - \widetilde{\beta}\right\rangle \|X\|_{\mathrm{F}}^2 \\
&\geqslant \frac{\eta\mu}{2}\|x^t - \widetilde{x}\|_{\mathcal{D}}^2 + \left(1 - \frac{\eta L}{2}\right)\left\{\|G^{t+1} - G^t\|_{\mathrm{F}}^2 + 2\left\|\left(\alpha^{t+1} - \alpha^t\right) 1_n{}^\top\right\|_{\mathrm{F}}^2 + \|(\beta^{t+1} - \beta^t)X\|_{\mathrm{F}}^2\right\}
\end{aligned}$$

where $\mu = \gamma_n$ and $L = \gamma_n + 9/2$.

Proof. Let $x^t = (G^t, \alpha^t, \beta^t)$ and $\widetilde{x} = (\widetilde{G}, \widetilde{\alpha}, \widetilde{\beta})$. Then

$$\begin{aligned}
&f(x^{t+1}) - f(\widetilde{x}) = f(x^{t+1}) - f(x^t) + f(x^t) - f(\widetilde{x}) \\
&\leqslant \left\langle \nabla f(x^t), x^{t+1} - x_t\right\rangle + \frac{L}{2}\|x^{t+1} - x^t\|_{\mathcal{D}}^2 + \left\langle \nabla f(x^t), x_t - \widetilde{x}\right\rangle - \frac{\mu}{2}\|x^t - \widetilde{x}\|_{\mathcal{D}}^2 \\
&\leqslant \left\langle \nabla f(x^t), x^{t+1} - \widetilde{x}\right\rangle + \frac{L}{2}\|x^{t+1} - x^t\|_{\mathcal{D}}^2 - \frac{\mu}{2}\|x^t - \widetilde{x}\|_{\mathcal{D}}^2 \\
&= \left\langle \nabla_G f(G^t, \alpha^t, \beta^t), G^{t+1} - \widetilde{G}\right\rangle + \left\langle \nabla_\alpha f(G^t, \alpha^t, \beta^t), \alpha^{t+1} - \widetilde{\alpha}\right\rangle + \left\langle \nabla_\beta f(G^t, \alpha^t, \beta^t), \beta^{t+1} - \widetilde{\beta}\right\rangle \\
&\quad + \frac{L}{2}\|x^{t+1} - x^t\|_{\mathcal{D}}^2 - \frac{\mu}{2}\|x^t - \widetilde{x}\|_{\mathcal{D}}^2.
\end{aligned}$$

Notice that $\widetilde{G}^{t+1} = G^t - \eta_G \frac{\partial f}{\partial G}|_{G=G^t}$ and G^{t+1} is the projection of $\widetilde{G}^{t+1}$ to the convex set $\{G | GJ = G, G \in \mathbb{S}_+, \max_{i,j}\|G_{ij}\| \leqslant M_1\}$. Therefore by Lemma A.5,

$$\left\langle G^{t+1} - \widetilde{G}^{t+1}, G^{t+1} - \widetilde{G}\right\rangle \leqslant 0$$

which implies that

$$\begin{aligned}
\left\langle \frac{\partial f}{\partial G}|_{G=G^t}, G^{t+1} - \widetilde{G}\right\rangle &\leqslant \frac{1}{\eta_G}\left\langle G^t - G^{t+1}, G^{t+1} - \widetilde{G}\right\rangle \\
&= \frac{1}{\eta_G}\left\langle G^t - G^{t+1}, G^t - \widetilde{G}\right\rangle - \frac{1}{\eta_G}\|G^t - G^{t+1}\|_{\mathrm{F}}^2.
\end{aligned}$$

Similar argument will yield

$$\Big\langle \frac{\partial f}{\partial \alpha}|_{\alpha=\alpha^{t+1}}, \alpha^{t+1}-\widetilde{\alpha}\Big\rangle \leqslant \frac{1}{\eta_\alpha}\Big\langle \alpha^t-\alpha^{t+1}, \alpha^t-\widetilde{\alpha}\Big\rangle - \frac{1}{\eta_\alpha}\|\alpha^t-\alpha^{t+1}\|^2,$$
$$\Big\langle \frac{\partial f}{\partial \beta}|_{\beta=\beta^{t+1}}, \beta^{t+1}-\widetilde{\beta}\Big\rangle \leqslant \frac{1}{\eta_\beta}\Big\langle \beta^t-\beta^{t+1}, \beta^t-\widetilde{\beta}\Big\rangle - \frac{1}{\eta_\beta}\|\beta^t-\beta^{t+1}\|^2.$$

Also notice that $f(x^{t+1})-f(\widetilde{x})\geqslant 0$, therefore

$$\begin{aligned}0 \leqslant \eta(f(x^{t+1})-f(\widetilde{x})) &\leqslant \big\langle G^t-G^{t+1}, G^t-\widetilde{G}\big\rangle + 2n\big\langle \alpha^t-\alpha^{t+1}, \alpha^t-\widetilde{\alpha}\big\rangle \\ &+ \|X\|_{\mathrm{F}}^2\big\langle \beta^t-\beta^{t+1}\big\rangle - \|x^t-x^{t+1}\|_{\mathcal{D}}^2 + \frac{\eta L}{2}\|x^t-x^{t+1}\|_{\mathcal{D}}^2 - \frac{\eta\mu}{2}\|x^t-\widetilde{x}\|_{\mathcal{D}}^2.\end{aligned}$$

This completes the proof. □

A.5.2 Proof of the theorem

Let $x^t=(G^t,\alpha^t,\beta^t)$, $\widetilde{x}=(\widetilde{G},\widetilde{\alpha},\widetilde{\beta})$. By definition,

$$\|x^{t+1}-\widetilde{x}\|_{\mathcal{D}}^2 = \|G^{t+1}-\widetilde{G}\|_{\mathrm{F}}^2 + 2\|\left(\alpha^{t+1}-\alpha^t\right)1_n{}^\top\|_{\mathrm{F}}^2 + \|(\beta^{t+1}-\widetilde{\beta})X\|_{\mathrm{F}}^2.$$

Notice that for each component, the error can be decomposed as (with G as an example),

$$\|G^{t+1}-\widetilde{G}\|_{\mathrm{F}}^2 = \|G^t-\widetilde{G}\|_{\mathrm{F}}^2 - 2\big\langle G^t-G^{t+1}, G^t-\widetilde{G}\big\rangle + \|G^{t+1}-G^t\|_{\mathrm{F}}^2.$$

Summing up these equations leads to

$$\begin{aligned}\|x^{t+1}-\widetilde{x}\|_{\mathcal{D}}^2 = \|x^t-\widetilde{x}\|_{\mathcal{D}}^2 &\\ - 2\Big\{\big\langle G^t-G^{t+1}, G^t-\widetilde{G}\big\rangle + 2n\big\langle \alpha^t-\alpha^{t+1}, \alpha^t-\widetilde{\alpha}\big\rangle + \|X\|_{\mathrm{F}}^2\big\langle \beta^t-\beta^{t+1}, \beta^t-\widetilde{\beta}\big\rangle\Big\} &\\ + \Big\{\|G^{t+1}-G^t\|_{\mathrm{F}}^2 + 2\|\left(\alpha^{t+1}-\alpha^t\right)1_n{}^\top\|_{\mathrm{F}}^2 + \|(\beta^{t+1}-\beta^t)X\|_{\mathrm{F}}^2\Big\}.&\end{aligned}$$

By Lemma A.6,

$$\|x^{t+1}-\widetilde{x}\|_{\mathcal{D}}^2 \leqslant (1-\eta\mu)\|x^t-\widetilde{x}\|_{\mathcal{D}}^2 - (1-\eta L)\|x^t-x^{t+1}\|_{\mathcal{D}}^2.$$

Then for any $\eta\leqslant 1/L$,

$$\|x^{t+1}-\widetilde{x}\|_{\mathcal{D}}^2 \leqslant (1-\eta\mu)\|x^t-\widetilde{x}\|_{\mathcal{D}}^2.$$

By Lemma 8.5, and repeatedly using the inequality $(a+b)^2\leqslant 2(a^2+b^2)$,

$$\begin{aligned}e_t &\leqslant \frac{\kappa_{Z_\star}^2}{2(\sqrt{2}-1)}\|Z^t(Z^t)^\top - Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2 \\ &\leqslant \frac{\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\left(\|Z^t(Z^t)^\top - G^t\|_{\mathrm{F}}^2 + \|G^t-Z_\star Z_\star^\top\|_{\mathrm{F}}^2\right) + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2 \\ &\leqslant \frac{2\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\|G^t-Z_\star Z_\star^\top\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2 \\ &\leqslant \frac{4\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\left(\|G^t-G_\star\|_{\mathrm{F}}^2 + \|G_\star-Z_\star Z_\star^\top\|_{\mathrm{F}}^2\right) + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2 \\ &\leqslant \frac{4\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\|G^t-G_\star\|_{\mathrm{F}}^2 + \frac{4\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\|\bar{G}_k\|_{\mathrm{F}}^2 + 2\|\Delta_{\alpha^t}1_n{}^\top\|_{\mathrm{F}}^2 + \|\Delta_{\beta^t}X\|_{\mathrm{F}}^2.\end{aligned}$$

By the definition of $\|\cdot\|_{\mathcal{D}}$, we further have

$$\begin{aligned}
e_t &\leqslant \frac{4\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\left(\|x^t - x_\star\|_{\mathcal{D}}^2 + \|\bar{G}_k\|_{\mathrm{F}}^2\right) \leqslant \frac{4\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\left(2\|x^t - \widetilde{x}\|_{\mathcal{D}}^2 + 2\|\widetilde{x} - x_\star\|_{\mathcal{D}}^2 + \|\bar{G}_k\|_{\mathrm{F}}^2\right) \\
&\leqslant \frac{4\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\left(2(1-\eta\gamma_n)^t\|x^0 - \widetilde{x}\|_{\mathcal{D}}^2 + 2\|\widetilde{x} - x_\star\|_{\mathcal{D}}^2 + \|\bar{G}_k\|_{\mathrm{F}}^2\right) \\
&\leqslant \frac{4\kappa_{Z_\star}^2}{(\sqrt{2}-1)}\left(4(1-\eta\gamma_n)^t\|x^0 - x_\star\|_{\mathcal{D}}^2 + 4(1-\eta\gamma_n)^t\|\widetilde{x} - x_\star\|_{\mathcal{D}}^2 + 2\|\widetilde{x} - x_\star\|_{\mathcal{D}}^2 + \|\bar{G}_k\|_{\mathrm{F}}^2\right).
\end{aligned}$$

According to Theorem A.2, there exists constant $C_0 > 0$ such that

$$\|\widetilde{x} - x_\star\|_{\mathcal{D}}^2 \leqslant C_0\left(e^{2M_1}\lambda_n^2 k + \frac{\|\bar{G}_k\|_*^2}{k}\right).$$

Therefore, $e_t \leqslant C_1\kappa_{Z_\star}^2\left((1-\eta\gamma_n)^t\|x^0 - x_\star\|_{\mathcal{D}}^2 + e^{2M_1}\lambda_n^2 k + \|\bar{G}_k\|_*^2/k + \|\bar{G}_k\|_{\mathrm{F}}^2\right)$. Since $x^0 = 0$,

$$\begin{aligned}
e_t &\leqslant C_1\kappa_{Z_\star}^2\left((1-\eta\gamma_n)^t\|x_\star\|_{\mathcal{D}}^2 + e^{2M_1}\lambda_n^2 k + \|\bar{G}_k\|_*^2/k + \|\bar{G}_k\|_{\mathrm{F}}^2\right) \\
&\leqslant \frac{c_1^2}{\kappa_{Z_\star}^4 e^{2M_1}}\|Z_\star\|_{\mathrm{op}}^4 \times \frac{C_1\kappa_{Z_\star}^6 e^{2M_1}}{c_1^2\|Z_\star\|_{\mathrm{op}}^4}\left((1-\eta\gamma_n)^t\|x_\star\|_{\mathcal{D}}^2 + e^{2M_1}\lambda_n^2 k + \|\bar{G}_k\|_*^2/k + \|\bar{G}_k\|_{\mathrm{F}}^2\right).
\end{aligned}$$

Under our assumptions, there exists some sufficiently large constant C_2 such that

$$\|Z_\star\|_{\mathrm{op}}^4 \geqslant C_2\kappa_{Z_\star}^6 e^{2M_1}\max\left\{e^{2M_1}\lambda_n^2 k,\ \|\bar{G}_k\|_*^2/k,\ \|\bar{G}_k\|_{\mathrm{F}}^2\right\}.$$

Therefore,

$$e_t \leqslant \frac{c_1^2}{\kappa_{Z_\star}^4 e^{2M_1}}\|Z_\star\|_{\mathrm{op}}^4 \times \left(\frac{C_1 e^{2M_1}\kappa_{Z_\star}^6\|\Theta_\star\|_{\mathrm{F}}^2}{c_1^2\tau^2\|Z_\star\|_{\mathrm{op}}^4}(1-\eta\gamma_n)^t + \frac{3C_1}{c_1^2 C_2}\right).$$

Choose large enough $C_2 > 6C_1/c_1^2$, then

$$e_t \leqslant \frac{c_1^2}{\kappa_{Z_\star}^4 e^{2M_1}}\|Z_\star\|_{\mathrm{op}}^4 \times \left(\frac{C e^{2M_1}\kappa_{Z_\star}^6\|x_\star\|_{\mathcal{D}}^2}{c_1^2\|Z_\star\|_{\mathrm{op}}^4}(1-\eta\gamma_n)^t + \frac{1}{2}\right).$$

Therefore, $e_t \leqslant \frac{c_1^2\tau^2}{\kappa^4}\|Z_\star\|_{\mathrm{op}}^4$ when

$$\frac{C_1 e^{2M_1}\kappa_{Z_\star}^6\|x_\star\|_{\mathcal{D}}^2}{c_1^2\|G_\star\|_{\mathrm{op}}^2}(1-\eta\gamma_n)^t \leqslant \frac{1}{2}.$$

By Lemma A.1, $\|G_\star\|_{\mathrm{op}}^2 \geqslant c\|G_\star\|_{\mathrm{F}}^2/k$. Therefore, it suffices to have

$$t \geqslant \log\left(k\frac{\|x_\star\|_{\mathcal{D}}^2}{\|G_\star\|_{\mathrm{F}}^2}\frac{2C_1 e^{2M_1}\kappa_{Z_\star}^6}{c_1^2 c}\right)\left(\log\left(\frac{1}{1-\eta\gamma_n}\right)\right)^{-1}.$$

This completes the proof.

B A Method for multiple edge covariates

In this appendix, we discuss the issue of fitting an inner-product model with multiple edge covariates.

B.1 Method

Suppose there are p different covariates, then one can extend model (1) to

$$A_{ij} = A_{ji} \overset{ind.}{\sim} \text{Bernoulli}(P_{ij}), \quad \text{with}$$
$$\text{logit}(P_{ij}) = \Theta_{ij} = \alpha_i + \alpha_j + \sum_{\ell=1}^{p} \beta_\ell X_{ij}^{(\ell)} + z_i^\top z_j,$$

where for $\ell = 1, \ldots, p$, the $n \times n$ symmetric matrix $X^{(\ell)} = (X_{ij}^{(\ell)})$ collects the observed values of the ℓ-th edge covariate, and $\beta = (\beta_1, \cdots, \beta_p)^T$ are the coefficients. For fixed tuning parameters λ_n^G and λ_n^X, our model fitting scheme solves the following convex program:

$$\begin{aligned} \min_{\alpha,\beta,G} \quad & -\sum_{i,j} \Big\{ A_{ij}\Theta_{ij} + \log\Big(1 - \sigma(\Theta_{ij})\Big)\Big\} + \lambda_n^G \,\text{Tr}(G) + \lambda_n^X \sum_{\ell=1}^{p} |\beta_\ell| \\ \text{subject to} \quad & \Theta = \alpha 1_n{}^\top + 1_n \alpha^\top + \sum_{\ell=1}^{p} \beta_\ell X^{(\ell)} + G, \;\; GJ = G, \;\; G \in \mathbb{S}_+^n, \quad -M_1 \leqslant \Theta_{ij} \leqslant -M_2. \end{aligned} \tag{50}$$

We propose an algorithm below which combines projected gradient descent and proximal gradient methods to solve (50).

Algorithm 4 A convex projected descent method for fitting model with multiple covariates.

1: **Input:** Adjacency matrix: A; covariate matrices: $X^{(1)}, \ldots, X^{(p)}$; latent space dimension: $k \geqslant 1$; tuning parameters: λ_n^G, λ_n^X; initial estimates: G^0, α^0, β^0; step sizes: $\eta_G, \eta_\alpha, \eta_\beta$; constraint sets: $\mathcal{C}_G, \mathcal{C}_\alpha, \mathcal{C}_\beta$.
   Output: $\widehat{G} = G^T$, $\widehat{\alpha} = \alpha^T$, $\widehat{\beta} = \beta^T$.
2: **for** $t = 0, 1, \cdots, T-1$ **do**
3: $\quad \widetilde{G}^{t+1} = G^t - \eta_G \nabla_G g(G, \alpha, \beta) = G^t + 2\eta_G \left(A - \sigma(\Theta^t) - \lambda_n^G I_n\right)$;
4: $\quad \widetilde{\alpha}^{t+1} = \alpha^t - \eta_\alpha \nabla_\alpha g(G, \alpha, \beta) = \alpha^t + 2\eta_\alpha (A - \sigma(\Theta^t)) 1_n$;
5: $\quad \widetilde{\beta}_\ell^{t+1} = \beta^t{}_\ell + \eta_\beta d_\ell^t$, where $d_\ell^t = \left\langle A - \sigma(\Theta^t), X^{(\ell)} \right\rangle$, for $1 \leqslant \ell \leqslant p$;
6: $\quad G^{t+1} = \mathcal{P}_{\mathcal{C}_G}(\widetilde{G}^{t+1})$, $\alpha^{t+1} = \mathcal{P}_{\mathcal{C}_\alpha}(\widetilde{\alpha}^{t+1})$, $\beta_\ell^{t+1} = S_{\eta_\beta \cdot \lambda_n^X}(\widetilde{\beta}_\ell^{t+1})$ for $1 \leqslant \ell \leqslant p$;
7: **end for**

In Step 6, $S_\delta(z) = \mathbf{1}_{\{|z|>\delta\}}(z - \text{sign}(z)\delta)$ represents the soft-thresholding operator. Note that the β-updates in steps 5 and 6 replicates the proximal gradient method for lasso [9]. Furthermore, we propose the following step sizes:

$$\eta_G = \eta, \quad \eta_\alpha = \eta/n, \quad \text{and} \quad \eta_\beta = \max_{1 \leqslant \ell \leqslant p} (\eta / \|X^{(\ell)}\|_{\text{F}}^2) \tag{51}$$

for some constant $\eta > 0$.

Tuning parameter selection To choose the optimal tuning parameters λ_n^G and λ_n^X, we suggest network cross-validation [51, 17] on a grid of these parameters. In particular, for a certain pair $(\lambda_n^G, \lambda_n^X)$ on the grid, we repeatedly partition the n nodes into I_1 and I_2 with $|I_1| = \lfloor n/2 \rfloor$ and $|I_2| = n - \lfloor n/2 \rfloor$. The edges $\{(i,j) : i \in I_1 \text{ or } j \in I_1\}$ are used for fitting the model (so that there are still n nodes, but the negative log-likelihood function only includes edges in this set), and the edges $\{(i,j) : i \in I_2 \text{ and } j \in I_2\}$ are used for testing. The optimal pair of tuning parameter is chosen so that the average mis-classification error for the testing edges over B random partitions is minimized. We recommend using a grid on the log scale around the center $(\bar{\lambda}_n^G, \bar{\lambda}_n^X) = (2\sqrt{n\widehat{p}}, n(n-1)\sqrt{\log n/n})$, where $\widehat{p} = \sum_{ij} A_{ij}/n^2$ is the average of all A's components. This validation scheme could help us choose good tuning parameters $\widehat{\lambda}_n^G$ and $\widehat{\lambda}_n^X$, as well as the covariates to include in our model. To avoid bias in estimating the β's, we may re-fit a model with the chosen $\widehat{\lambda}_n^G$ and the included covariates, but without the L_1-penalty for the β's.

For a model without any covariate, there is a similar validation scheme to the one described in the preceding paragraph for choosing the optimal λ_n^G. Specifically, for a fixed λ_n^G, Algorithm 4 applied by removing Step 5 and the β-update in Step 6, and the optimal λ_n^G is one that minimizes the average testing mis-classification error over B random partitions in a λ_n^G sequence.

B.2 The lawyer data example

We now revisit the lawyer data example in Section 6.2. We label attributes practice, gender, office, and school as node covariate 1 to 4. Then we set edge covariate $X_{ij}^{(\ell)} = X_{ji}^{(\ell)} = 1$ if $i \neq j$ and the ith and the jth lawyers shared the same ℓth node covariate, and $X_{ij}^{(\ell)} = X_{ji}^{(\ell)} = 0$ otherwise. Applying Algorithm 4 with the aforementioned turning parameter selection scheme suggests that the first three covariates, i.e., indicators for attributes practice, gender, and office, should be included in the model. The resulting number of misclustered nodes is 9, a 25% improvement compared to the previously reported 12 misclustered nodes when not using any covariate. This error rate is slightly worse than including a single handpicked covariate, indicator for practice, but it enjoys the merit of not having to manually choose the covariates or tune the parameters. Finally, we also note that the error rate does not change if one use latent vectors derived from all non-zero eigenvalues of the fitted $\widehat{G}$.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43. ACM, 2005.

[2] D. Agarwal and B.-C. Chen. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 19–28. ACM, 2009.

[3] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[4] E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a

graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[5] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

[6] D. Asta and C. R. Shalizi. Geometric network comparison. *arXiv preprint arXiv:1411.1350*, 2014.

[7] P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[8] E. Bogomolny, O. Bohigas, and C. Schmit. Spectral properties of distance matrices. *Journal of Physics A: Mathematical and General*, 36(12):3595, 2003.

[9] S. Boyd and N. Parikh. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 123–231, 2013.

[10] S. Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 2014.

[11] S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[12] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[13] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[14] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[15] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[16] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[17] K. Chen and J. Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

[18] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[19] Y. Chen, X. Li, and J. Xu. Convexified modularity maximization for degree-corrected stochastic block models. *arXiv preprint arXiv:1512.08425*, 2015.

[20] X. Cheng and A. Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

[21] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145, 2002.

[22] M. A. Davenport, Y. Plan, E. Van Den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[23] R. L. Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

[24] N. El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[25] C. Gao, Y. Lu, H. H. Zhou, et al. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

[26] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.

[27] C. Gao, Y. Lu, Z. Ma, and H. H. Zhou. Optimal estimation and completion of matrices with biclustering structures. *Journal of Machine Learning Research*, 17(161):1–29, 2016.

[28] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[29] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

[30] M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[31] P. D. Hoff. Random effects models for network data. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*. Citeseer, 2003.

[32] P. D. Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the american Statistical association*, 100(469):286–295, 2005.

[33] P. D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in neural information processing systems*, pages 657–664, 2008.

[34] 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.

[35] J. Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

[36] J. Jin, Z. T. Ke, and S. Luo. Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*, 2017.

[37] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[38] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

[39] O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. *arXiv preprint arXiv:1507.04118*, 2015.

[40] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329, 2011.

[41] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

[42] P. N. Krivitsky and M. S. Handcock. Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software*, 24(i05), 2008.

[43] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213, 2009.

[44] E. Lazega. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press on Demand, 2001.

[45] J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[46] Z. Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[47] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

[48] M. Mézard, G. Parisi, and A. Zee. Spectra of Euclidean random matrices. *Nuclear Physics B*, 559(3):689–701, 1999.

[49] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[50] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

[51] A. B. Owen and P. O. Perry. Bi-cross-validation of the svd and the nonnegative matrix factorization. *The annals of applied statistics*, pages 564–594, 2009.

[52] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[53] I. J. Schoenberg. On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space. *Annals of Mathematics*, pages 787–793, 1937.

[54] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[55] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[56] D. L. Sussman, M. Tang, and C. E. Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57, 2014.

[57] M. Tang, D. L. Sussman, and C. E. Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, 2013.

[58] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

[59] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[60] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[61] P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

[62] Y.-J. Wu, E. Levina, and J. Zhu. Generalized linear models with low rank effects for network data. *arXiv preprint arXiv:1705.06772*, 2017.

[63] S. J. Young and E. R. Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

[64] Y. Zhang, E. Levina, and J. Zhu. Detecting overlapping communities in networks using spectral methods. *arXiv preprint arXiv:1412.3432*, 2014.

[65] Y. Zhang, E. Levina, and J. Zhu. Community detection in networks with node features. *arXiv preprint arXiv:1509.01173*, 2015.

[66] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *stat*, 1050:23, 2016.