Non-Convex Joint Community Detection and Group Synchronization via Generalized Power Method

Sijin Chen
Princeton University

Xiwei Cheng
CUHK

Anthony Man-Cho So
CUHK

Abstract

This paper proposes a Generalized Power Method (GPM) to simultaneously solve the joint problem of community detection and group synchronization in a direct non-convex manner, in contrast to the existing method of semidefinite programming (SDP). Under a natural extension of stochastic block model (SBM), our theoretical analysis proves that the proposed algorithm is able to exactly recover the ground truth in $O(n \log^2 n)$ time for problems of size n, sharply outperforming the $O(n^{3.5})$ runtime of SDP. Moreover, we give a lower bound of model parameters as a sufficient condition for the exact recovery of GPM. The new bound breaches the information-theoretic limit for pure community detection under SBM, thus demonstrating the superiority of our simultaneous optimization algorithm over any two-stage method that performs the two tasks in succession. We also conduct numerical experiments on GPM and SDP to corroborate our theoretical analysis.

1 INTRODUCTION

Community detection methods typically make use of the edge connectedness information of an observation network to infer the underlying clustering of the nodes or agents participating in the network (Abbe, 2018; Yun and Proutiere, 2014; Amini and Levina, 2018; Gao et al., 2017). However, if additional information about the nodes besides the edge connectedness is available, chances are that the recovery results of the clustering can outperform the pure community detection methods, even its information-theoretic limit (Abbe and Sandon, 2015), by carefully exploiting this extra information (Weng and Feng, 2016; Binkiewicz et al., 2017; Zhang et al., 2016). Based on this fact, one may proceed to wonder whether a simultaneous recovery of both the clustering and the additional nodal features can be achieved efficiently, or, more efficiently than the trivial two-stage method which first infers a community structure and then recovers the nodal features assuming this community structure.

In this paper, we answer this question *affirmatively* as for the scenario where the additional information comes from the node-wise relative measurements belonging to a matrix group $\mathcal{G}$, which arises in group synchronization problems (Arie-Nachimson et al., 2012; Boumal, 2016; Bandeira et al., 2016; Liu et al., 2017, 2023). Such modeling of community detection with additional node-wise measurement finds applications in a number of practical problems, for example, the 2D class averaging of cryo-electron microscopy single-particle reconstruction (Frank, 2006; Singer et al., 2011; Zhao and Singer, 2014), and simultaneous mapping and clustering of 3D object volumes in computer vision (Bajaj et al., 2018).

Recently, Fan et al. (2022) looked into this problem and defined a model with its optimization problem, which we inherit with necessary adjustments and generalizations. The model naturally extends the celebrated stochastic block model (SBM) (Abbe and Sandon, 2015) for community detection. Assume that there are n agents in a network partitioned into K communities, and each agent i corresponds to a group element $g_i \in \mathcal{G}$. With probability p, we obtain the relative group measurement $g_i g_j^{-1} \in \mathcal{G}$ for two agents i, j belonging to the same cluster; with probability q, we obtain an observation noise g uniformly sampled from $\mathcal{G}$ for two agents i, j falling in different clusters – a mechanism also named the out-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume TBD. Copyright 2024 by the author(s).

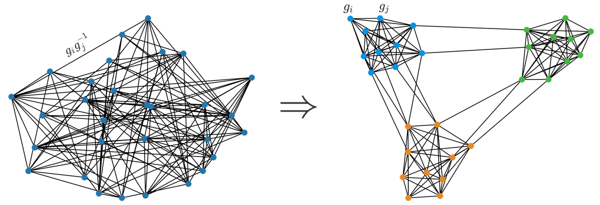


Figure 1: Illustration of the joint optimization problem of community detection and group synchronization. Left: we observe a network of $n = 30$ agents falling in $K = 3$ equal-sized communities, with relative measurement $g_ig_j^{-1} \in \mathcal{G}$ between any connected pair of nodes. Right: the target is to recover both the underlying cluster and the group element g_i of each agent simultaneously.

lier noise model (Singer, 2011; Fan and Zhao, 2019). Given these observations, we aim to recover both the underlying clustering and the corresponding group elements of all the agents. Figure 1 illustrates the problem settings and the target of recovery.

The above can be formally stated as a non-convex optimization problem (Fan et al., 2022) that jointly involves the clustering variables and the rotational group variables. As a result of this joint formulation, through solving this problem, the clustering and synchronization tasks are performed at the same time. Due to the computational hardness of the problem, Fan et al. (2022) designed a semidefinite relaxation for the original problem, and then solved the semidefinite program (SDP) with numerical methods. By exploiting the mutual rotational measurements to recover the underlying community and rotation of each agent simultaneously, it enjoys both theoretical and experimental advantages over the state-of-the-art pure community detection methods, and hence the naive two-stage method.

However, a major challenge of applying SDP to large-scale practice is still its time complexity. From the theory side, the standard interior point method for SDP typically takes $O(n^{3.5})$ time to find an optimal solution, which is rather computationally expensive. This is evidenced by our numerical experiments (see Section 6), where it already takes an unacceptable time to solve the problem of $n \approx 200$ with standard MATLAB SDP implementation. Moreover, the theory developed in Fan et al. (2022) still leaves several open questions. For example, it only takes care of the $\mathcal{G} = \mathcal{SO}(d)$ situation, and rigorous analysis is only provided for $K = 2$. These issues hinder the reliability of SDP when applied to general cases.

1.1 Contributions

Different from the methodology of convex relaxation adopted by Fan et al. (2022), this paper proposes an iterative *Generalized Power Method* (GPM) to directly tackle the non-convex optimization problem of simultaneous community detection with group synchronization, and provides a theoretical guarantee for its linear convergence to the optimal solution under certain conditions. Our contributions are threefold.

- **Significant runtime boost, no discount on boundary.** Our algorithm sharply reduces the time complexity to $O(n \log^2 n)$ from $O(n^{3.5})$ of SDP without any compensation on the lower bound for model parameters. Our numerical studies also indicate a significantly boosted runtime with no discount on the phase transition boundary. The iterative algorithm is also structurally simple and practically convenient to implement.
- **Broader coverage of theory.** In this paper, theoretical guarantees are provided for both rotational ($\mathcal{G} = \mathcal{SO}(d)$) and orthogonal ($\mathcal{G} = \mathcal{O}(d)$) cases, and the number of clusters is allowed to be $\Theta(1)$. These generalize the scenario considered by Fan et al. (2022) of rotational synchronization with only 2 clusters.
- **Outperforming any pure community detection method.** We also remark that the conditions for linear convergence in this paper are able to break the information-theoretic lower bound $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{K}$ for pure community detection (Abbe and Sandon, 2015), thus demonstrating the superiority of our joint method over any naive two-stage approach that invokes a pure community detection method.

1.2 Organization

The rest of this paper is organized as follows. Section 2 briefly introduces the preliminaries about community detection and group synchronization, which are important to the presentation and analysis of our joint problem. Then, Section 3 presents a formal definition of our probabilistic model and formulates the non-convex optimization problem of simultaneous community detection with group synchronization. Our non-convex methodology is elaborated in Section 4, followed by a detailed statement of our main theoretical results in Section 5. Their proofs are deferred to the Appendix. Section 6 presents the numerical results from computer simulations. Finally, we make some discussions and conclude this paper in Section 7.

1.3 Notation

We use $\boldsymbol{X}^\top$ to denote the transpose of a matrix $\boldsymbol{X}$. If a matrix $\boldsymbol{X}$ has a $d\times d$-block structure, then we use $\boldsymbol{X}_{ij}$ to refer to the (i,j)-th $d\times d$ block of $\boldsymbol{X}$ and $\boldsymbol{X}_{i\times}$ to refer to the i-th block row of $\boldsymbol{X}$. The $n\times n$ identity matrix is denoted by $\boldsymbol{I}_n$, the $n\times m$ all-one matrix is denoted by $\mathbf{1}_{n\times m}$, and the $n\times m$ all-zero matrix is denoted by $\mathbf{0}_{n\times m}$. We use $\otimes$ and $\odot$ to represent the Kronecker product and Hadamard (elementwise) product of two matrices with conforming shapes.

For matrix $\boldsymbol{X}\in\mathbb{R}^{n\times n}$ and integer $k\le n$, $\sigma_k(\boldsymbol{X})$ is the k-th largest singular value of $\boldsymbol{X}$. $\|\boldsymbol{X}\|:=\sigma_1(\boldsymbol{X})$, $\|\boldsymbol{X}\|_F^2:=\operatorname{tr}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)$, and $\|\boldsymbol{X}\|_*:=\sum_{k=1}^n\sigma_k(\boldsymbol{X})$ respectively represents the operator norm, Frobenius norm, and nuclear norm of $\boldsymbol{X}$. We denote $\langle\boldsymbol{X},\boldsymbol{Y}\rangle:=\operatorname{tr}\left(\boldsymbol{X}^\top\boldsymbol{Y}\right)$ the (Frobenius) inner product of matrices $\boldsymbol{X},\boldsymbol{Y}$. $\lambda_k(\boldsymbol{X})$ denotes its k-th largest (real) eigenvalue for symmetric $\boldsymbol{X}$. For symmetric $\boldsymbol{X}\in\mathbb{R}^{n\times n}$, we define $\boldsymbol{U}:=\operatorname{eigs}_{[k:l]}(\boldsymbol{X})$ such that the columns of $\boldsymbol{U}$ consist of the unit eigenvectors of $\boldsymbol{X}$ associated to eigenvalues $\lambda_k(\boldsymbol{X}),\lambda_{k+1}(\boldsymbol{X}),...,\lambda_l(\boldsymbol{X})$. we simply write $\boldsymbol{U}=\operatorname{eigs}_l(\boldsymbol{X})$ if $k=1$.

For any integer $n\ge 1$ we define $[n]:=\{1,2,...,n\}$.

2 PRELIMINARIES

In this section, we quickly go through two classical problems, each of which enjoys an independent and rich line of works: *community detection* and *group synchronization*. Building up the fundamentals of these two problems is essential for the development of our main content.

Community Detection

The main problem to be studied in this paper has a strong correlation to *community detection* problems under the symmetric stochastic block model (SBM) (Abbe and Sandon, 2015) with parameters (n,p,q,K). Assume that n agents in a network fall in K underlying communities of equal size $m=n/K$ (balanced clustering). Then, SBM(n,p,q,K) generates a random undirected graph G such that every two nodes are connected by an edge with probability p if they belong to the same cluster, and with probability q otherwise. We assume without loss of generality that a node is always connected to itself in SBM(n,p,q).

Clustering functions and clustering matrices (Wang et al., 2021) are defined to formally represent the community structure of the nodes. We denote $C:[n]\to[K]$ the clustering function that maps node i to cluster $C(i)$ where it belongs. Conversely, we define $\mathcal{I}_j:=\{i\in[n]\mid C(i)=j\}$ to be the set of nodes in cluster j. We can one-hot encode the clustering function in a clustering matrix $\boldsymbol{H}\in\{0,1\}^{n\times K}$, such that $\boldsymbol{H}_{ij}=1$ if and only if $C(i)=j$. Furthermore, one can show that the following $\mathcal{H}$ is the set of all $n\times K$ clustering matrices

$$\mathcal{H}:=\{\boldsymbol{H}\in\{0,1\}^{n\times K}\;|\boldsymbol{H}\mathbf{1}_{K\times 1}=\mathbf{1}_{n\times 1},\\ \mathbf{1}_{1\times n}\boldsymbol{H}=m\mathbf{1}_{1\times K}\}.$$

Our cluster structure is invariant under permutations of the cluster numbering. As a result, we can define the equivalent class of an $\boldsymbol{H}\in\mathcal{H}$ to be $\{\boldsymbol{HP}\mid\boldsymbol{P}\in\mathcal{S}_K\}$, where $\mathcal{S}_K$ is the K-dimensional permutation group. Therefore, given a ground truth $\boldsymbol{H}^*$, the estimation error (Wang et al., 2021) of $\boldsymbol{H}\in\mathcal{H}$ is defined as

$$\epsilon(\boldsymbol{H})=\min_{\boldsymbol{P}\in\mathcal{S}_K}\|\boldsymbol{H}^*-\boldsymbol{HP}\|_F.$$

Now we discuss the computational cost of projecting onto $\mathcal{H}$. For arbitrary matrix $\boldsymbol{M}\in\mathbb{R}^{n\times K}$, we say

$$\Pi_{\mathcal{H}}(\boldsymbol{M}):=\operatorname*{argmin}_{\boldsymbol{H}\in\mathcal{H}}\|\boldsymbol{M}-\boldsymbol{H}\|_F \tag{1}$$

is the projection of $\boldsymbol{M}$ onto $\boldsymbol{H}$. As is pointed out by Wang et al. (2021), Problem 1 is equivalent to a *minimum-cost assignment problem* (MCAP) that can be tackled efficiently by existing algorithms (Tokuyama and Nakano, 1995). The following proposition by Wang et al. (2021) gives an upper bound on its time complexity.

Proposition 1 (Proposition 1, Wang et al. (2021))**.** *Problem 1 can be solved in $O(K^2 n\log n)$ time.*

When the parameters p and q are located in the logarithmic sparsity region of SBM(n,p,q,K), *i.e.*, $p=\frac{\alpha\log n}{n}$ and $\frac{\beta\log n}{n}$ where $\alpha,\beta=O(1)$, Abbe and Sandon (2015) derived that one can recover the underlying clustering if and only if

$$\sqrt{\alpha}-\sqrt{\beta}>\sqrt{K}$$

under SBM. This is the *information-theoretic limit* for the *pure* community detection problems.

Group Synchronization

Our main problem interacts with another nonconvex optimization problem named *group synchronization*, and our focus will be on the synchronization of orthogonal and rotational groups. Recall the

d-dimensional orthogonal group over $\mathbb{R}$ is

$$\mathcal{O}(d) := \left\{ \boldsymbol{Q} \in \mathbb{R}^{d\times d} \;\middle|\; \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_d \right\}$$

with the usual matrix multiplication as the group operation. Moreover, the d-dimensional rotational group $\mathcal{SO}(d)$, or special orthogonal group, consists of all $\mathcal{O}(d)$ matrices with determinant 1.

In the typical formulation of group synchronization (Boumal, 2016; Liu et al., 2017, 2023), there are n agents in a measurement network and the i-th agent corresponds to a group element $g_i \in \mathcal{G}$, where we specially consider $\mathcal{G} \in \{\mathcal{O}(d), \mathcal{SO}(d)\}$ in this paper. For group $\mathcal{G}$, we define a set of block matrices $\mathcal{G}^n \subset \mathbb{R}^{nd\times d}$ as follows:

$$\mathcal{G}^n := \left\{ \boldsymbol{X} \in \mathbb{R}^{nd\times d} \;\middle|\; \boldsymbol{X} = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix}, g_i \in \mathcal{G} \right\}.$$

Moreover, we define a block-diagonalization operator $\operatorname{bdiag}(\cdot)$ such that for $g_1, g_2, \cdots, g_n \in \mathcal{G}$,

$$\operatorname{bdiag}(\boldsymbol{X}) := \begin{pmatrix} g_1 & & & \\ & g_2 & & \\ & & \ddots & \\ & & & g_n \end{pmatrix}, \forall \boldsymbol{X} = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix}.$$

And we denote $\operatorname{bdiag}(\mathcal{G}^n) := \{\operatorname{bdiag}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{G}^n\}$.

For arbitrary matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, we define

$$\Pi_{\mathcal{O}(d)}(\boldsymbol{X}) := \operatorname*{argmin}_{\boldsymbol{Q}\in\mathcal{O}(d)} \|\boldsymbol{X} - \boldsymbol{Q}\|_F \tag{2}$$

as the projection of $\boldsymbol{X}$ onto $\mathcal{O}(d)$ (Arie-Nachimson et al., 2012; Ling, 2022; Liu et al., 2023). Problem 2 can be categorized into the *orthogonal Procrustes problem* (Gower and Dijksterhuis, 2004) that has a closed-form solution:

Proposition 2. *For any given $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, suppose that $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the SVD of $\boldsymbol{X}$. Then, $\Pi_{\mathcal{O}(d)}(\boldsymbol{X}) = \boldsymbol{U}\boldsymbol{V}^\top$, and $\max_{\boldsymbol{Q}\in\mathcal{O}(d)} \langle \boldsymbol{X}, \boldsymbol{Q}\rangle = \operatorname{tr}(\boldsymbol{\Sigma}) = \|\boldsymbol{X}\|_*$.*

3 PROBLEM FORMULATION

In this section, we define our probabilistic model and introduce the concerning optimization problem proposed by Fan et al. (2022). We then reformulate an equivalent variant of this problem to fit our nonconvex methodology, and build up some useful infrastructures such as rounding method, orthogonal projection, and error metric.

3.1 The Probabilistic Model and the Joint Optimization Problem

We consider a *stochastic group block model* (SGBM) with parameters $(n, p, q, K, d, \mathcal{G})$, and formulate the corresponding joint optimization problem. Assume that n agents in a network fall in K underlying communities of equal size $m = n/K$, and each agent i corresponds to a group element $\boldsymbol{R}_i^* \in \mathcal{G}$, where $\mathcal{G} \in \{\mathcal{O}(d), \mathcal{SO}(d)\}$. All the group elements constitute a block matrix $\boldsymbol{R}^* \in \mathcal{G}^n$. Denote C^* the clustering function, $\mathcal{I}_k^*$ the set of agents $\{i : C^*(i) = k\}$, and $\boldsymbol{H}^* \in \mathcal{H}$ the clustering matrix.

$\mathrm{SGBM}(n, p, q, K, d, \mathcal{G})$ firstly generates a random undirected graph $G = (V, E)$ under $\mathrm{SBM}(n, p, q, K)$. Then, it generates an observation matrix $\boldsymbol{A} \in \mathbb{R}^{nd\times nd}$, whose (i, j)-th block is defined by the following process:

1. if $\{i, j\} \notin E$, then $\boldsymbol{A}_{ij} = \boldsymbol{0}_{d\times d}$; otherwise,
2. if $C^*(i) = C^*(j)$, then $\boldsymbol{A}_{ij} = \boldsymbol{R}_i^* \boldsymbol{R}_j^{*\top}$;
3. if $C^*(i) \neq C^*(j)$,
   (a) if $i < j$, then $\boldsymbol{A}_{ij} = \boldsymbol{R}_{ij} \sim \mathrm{Unif}(\mathcal{G})$ which is the uniform distribution over $\mathcal{G}$ with respect to the Haar measure;
   (b) if $i > j$, then $\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji}^\top$.

Given the observation matrix $\boldsymbol{A}$, Fan et al. (2022) introduced Problem Joint-$\mathcal{G}$ that aims to jointly recover the community structure $\boldsymbol{H} \in \mathcal{H}$ and the group elements $\boldsymbol{R} \in \mathcal{G}^n$:

$$\max_{\substack{\boldsymbol{R}\in\mathcal{G}^n \\ \boldsymbol{H}\in\mathcal{H}}} \sum_{k=1}^{K} \sum_{i,j\in\mathcal{I}_k} \left\langle \boldsymbol{A}_{ij}, \boldsymbol{R}_i \boldsymbol{R}_j^\top \right\rangle. \tag{Joint-$\mathcal{G}$}$$

3.2 Reformulation of the Problem

The non-convexity of Problem Joint-$\mathcal{G}$ makes it computationally intractable without proper reformulations. One can integrate the group variables $\boldsymbol{R}$ and clustering variables $\boldsymbol{H}$ into a block matrix $\boldsymbol{M}(\boldsymbol{R}, \boldsymbol{H}) \in \mathbb{R}^{nd\times nd}$, such that its (i, j)-block

$$\boldsymbol{M}_{ij} = \begin{cases} \boldsymbol{R}_i \boldsymbol{R}_j^\top, & \text{if } C(i) = C(j), \\ \boldsymbol{0}, & \text{otherwise.} \end{cases} \tag{3}$$

Now, Problem Joint-$\mathcal{G}$ is equivalently cast to

$$\max_{\substack{\boldsymbol{R}\in\mathcal{G}^n \\ \boldsymbol{H}\in\mathcal{H}}} \langle \boldsymbol{A}, \boldsymbol{M}\rangle. \tag{4}$$

A blessing of introducing $\boldsymbol{M}$ into the problem is that $\boldsymbol{M}$ actually admits a low-rank decomposition, which paves the way for further linear algebra manipulations.

Proposition 3. *For any $\boldsymbol{R} \in \mathcal{G}^n$ and $\boldsymbol{H} \in \mathcal{H}$, let*

$$\boldsymbol{V} = (\mathbf{1}_{1\times K} \otimes \boldsymbol{R}) \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d}) \in \mathbb{R}^{nd\times Kd},$$

and $\boldsymbol{M} \in \mathbb{R}^{nd\times nd}$ as defined by (3). Then we have $\boldsymbol{M} = \boldsymbol{V}\boldsymbol{V}^\top$. Moreover, $\boldsymbol{V}^\top \boldsymbol{V} = m\boldsymbol{I}_{Kd}$ and hence $\frac{1}{\sqrt{m}}\boldsymbol{V}$ is orthonormal.

By Proposition 3, Problem 4 is reformulated as a quadratic program subject to non-convex constraints

$$\begin{aligned}\max_{\substack{\boldsymbol{R}\in\mathcal{G}^n\\ \boldsymbol{H}\in\mathcal{H}}} \langle \boldsymbol{A}, \boldsymbol{M}\rangle &= \max_{\boldsymbol{V}\in\mathcal{E}} \left\langle \boldsymbol{A}, \boldsymbol{V}\boldsymbol{V}^\top \right\rangle \\ &= \max_{\boldsymbol{V}\in\mathcal{E}} \operatorname{tr}(\boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}), \qquad (\mathsf{JointQP}\text{-}\mathcal{G})\end{aligned}$$

where the feasible region

$$\mathcal{E} := \{(\mathbf{1}_{1\times K} \otimes \boldsymbol{R}) \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d}) \mid \boldsymbol{R} \in \mathcal{G}^n, \boldsymbol{H} \in \mathcal{H}\}$$

is non-convex. Since $\mathcal{SO}(d)$ is a subgroup of $\mathcal{O}(d)$, Problem JointQP-$\mathcal{G}$ can be relaxed to

$$\max_{\boldsymbol{V}\in\mathcal{F}} \operatorname{tr}(\boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}), \qquad (\mathsf{JointQP}\text{-}\mathcal{O}(d))$$

where $\mathcal{F} := \{(\mathbf{1}_{1\times K} \otimes \boldsymbol{R}) \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d}) \,|\boldsymbol{R} \in \mathcal{O}(d)^n, \boldsymbol{H} \in \mathcal{H}\}$ is again a non-convex feasible region. This is the ultimate problem we will study in this paper.

3.3 Rounding and Projection

When $\mathcal{E} \neq \mathcal{F}$, it is necessary to introduce a rounding function $\mathcal{R}$ that maps a matrix $\boldsymbol{V} \in \mathcal{F}$ back to $\mathcal{E}$. The rounding function is defined blockwise:

$$\mathcal{R}(\boldsymbol{V})_{ij} = \det(\boldsymbol{V}_{ij})\boldsymbol{V}_{ij}, \ \forall i \in [n], j \in [K].$$

Thanks to this simple rounding function $\mathcal{R}$, the relaxation from $\mathcal{G}^n$ to $\mathcal{O}(d)^n$ is proved to be sufficiently tight. The proof is deferred to the Appendix.

For arbitrary $\boldsymbol{X} \in \mathbb{R}^{nd\times Kd}$, we say

$$\Pi_{\mathcal{F}}(\boldsymbol{X}) := \operatorname*{argmin}_{\boldsymbol{W}\in\mathcal{F}} \|\boldsymbol{W} - \boldsymbol{X}\|_F$$

is the projection of $\boldsymbol{X}$ onto $\mathcal{F}$. In Section 4, we will provide Algorithm 2 to efficiently compute this projection.

3.4 The Metric of Estimation Error

Consider an arbitrary $\boldsymbol{V} \in \mathcal{E}$. Similar to the permutation invariance mentioned in Section 2, any permutation of the clusters makes no difference under the settings of SGBM, since the generation of the observation matrix does not depend on the specific numbering of clusters. Formally, any $\boldsymbol{V}\boldsymbol{Q} \in \mathcal{E}$ would yield the same probabilistic distribution of observation as $\boldsymbol{V}$, where $\boldsymbol{Q} \in \mathcal{Q} := \{\boldsymbol{P} \otimes \boldsymbol{I}_d \mid \boldsymbol{P} \in \mathcal{S}_K\}$ and $\mathcal{S}_K$ is the permutation group on $[K]$, represented by $K \times K$ permutation matrices.

Moreover, right-multiplying any element of $\mathcal{G}$ commonly on each cluster neither affects the observation it induces. This is because, for any i, j belonging to the same cluster with group elements $\boldsymbol{R}_i, \boldsymbol{R}_j$, their relative measurement remains intact after a common right multiplication, *i.e.*

$$(\boldsymbol{R}_i\boldsymbol{U})(\boldsymbol{R}_j\boldsymbol{U})^\top = \boldsymbol{R}_i\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{R}_j^\top = \boldsymbol{R}_i\boldsymbol{R}_j^\top, \ \forall \boldsymbol{U} \in \mathcal{G}.$$

In block matrix language, for any $\boldsymbol{W} \in \operatorname{bdiag}\left(\mathcal{G}^K\right)$ where

$$\operatorname{bdiag}\left(\mathcal{G}^K\right) := \left\{ \begin{pmatrix} \boldsymbol{U}_1 & & & \\ & \boldsymbol{U}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{U}_K \end{pmatrix} \;\middle|\; \substack{\boldsymbol{U}_i\in\mathcal{G}\\ \forall i\in[n]} \right\},$$

$\boldsymbol{V}\boldsymbol{W} \in \mathcal{E}$ also yields the same probabilistic distribution of observation as $\boldsymbol{V}$.

We can unify both permutation invariance $\mathcal{Q}$ and orthogonality invariance $\operatorname{bdiag}\left(\mathcal{G}^K\right)$ by defining a group $\mathcal{P}_K(\mathcal{G})$ of $\mathbb{R}^{Kd\times Kd}$ matrices, such that $\boldsymbol{Q} \in \mathcal{P}_K(\mathcal{G})$ if and only if

$$\exists \boldsymbol{R} \in \mathcal{G}^K, \boldsymbol{P} \in \mathcal{S}_K : \boldsymbol{Q} = \operatorname{bdiag}(\boldsymbol{R})(\boldsymbol{P} \otimes \boldsymbol{I}_d).$$

Now, an equivalence class of $\boldsymbol{V}$ is given by the orbit of $\boldsymbol{V}$ under the $\mathcal{P}_K(\mathcal{G})$ actions, which inspires us to define an *orbit distance* for two points $\boldsymbol{V}_1, \boldsymbol{V}_2$.

Definition 1 (Orbit distance)**.** *For any $\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathcal{E}$,*

$$\operatorname{dist}_{\mathcal{G}}(\boldsymbol{V}_1, \boldsymbol{V}_2) := \min_{\boldsymbol{Q}\in\mathcal{P}_K(\mathcal{G})} \|\boldsymbol{V}_1 - \boldsymbol{V}_2\boldsymbol{Q}\|_F$$

is said to be the orbit distance between $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$.

It is easy to verify that $\operatorname{dist}_{\mathcal{G}}$ is a metric. In order to properly measure the estimation error of an arbitrary $\boldsymbol{V} \in \mathcal{E}$ and the ground truth $\boldsymbol{V}^*$, we should use the orbit distance to account for the invariances underlying the observation.

Definition 2 (Estimation error)**.** *For any $\boldsymbol{V}, \boldsymbol{V}^* \in \mathcal{E}$ where $\boldsymbol{V}^*$ is the ground truth,*

$$\epsilon_{\mathcal{G}}(\boldsymbol{V}) := \operatorname{dist}_{\mathcal{G}}(\boldsymbol{V}, \boldsymbol{V}^*) = \min_{\boldsymbol{Q}\in\mathcal{P}_K(\mathcal{G})} \|\boldsymbol{V} - \boldsymbol{V}^*\boldsymbol{Q}\|_F$$

is said to be the estimation error of $\boldsymbol{V}$.

4 NONCONVEX METHODOLOGY

We now propose a *Generalized Power Method* (GPM) to tackle Problem JointQP-$\mathcal{O}(d)$; cf. Boumal

(2016); Liu et al. (2017, 2023); Wang et al. (2021, 2022a,b):

Algorithm 1 GPM

```
**Input:** the observation matrix $\boldsymbol{A}$, an initial point $\boldsymbol{V}^0$.
**for** $t = 0, 1, 2, \ldots, T-1$ **do**
    $\boldsymbol{V}^{t+1} \leftarrow \Pi_{\mathcal{F}}(\boldsymbol{A}\boldsymbol{V}^t)$
**end for**
**if** $\mathcal{G} = \mathcal{SO}(d)$ **then**
    $\boldsymbol{V}^T \leftarrow \mathcal{R}(\boldsymbol{V}^T)$
**end if**
**Return:** $\boldsymbol{V}^T$
```

The structure of Algorithm 1 is concise. It iteratively refines the initial guess $\boldsymbol{V}^0$ by simply taking matrix multiplication and projection onto the relaxed feasible region $\mathcal{F}$. The algorithm directly outputs the last iteration in the orthogonal case $\mathcal{G} = \mathcal{O}(d)$, and rounds the last iteration to feasibility in the rotational case $\mathcal{G} = \mathcal{SO}(d)$. GPM does not introduce any hyperparameters and is convenient to deploy in practice.

To analyze the time complexity of Algorithm 1, we first note that due to the logarithmic sparsity assumption, every block row of the observation matrix $\boldsymbol{A}$ has $O(\log n)$ non-zero blocks. It follows that $\boldsymbol{AV}$ can be computed in $O(n \log n)$ time by exploiting the sparsity of $\boldsymbol{A}$. The only black box in Algorithm 1 – until now – is the computation of the projection operator $\Pi_{\mathcal{F}}$. Here we unveil it by providing an efficient algorithm to solve the projection. To this end, define a mapping $\mu : \mathbb{R}^{nd \times Kd} \to \mathbb{R}^{n \times K}$ that computes the blockwise nuclear norm

$$\mu(\boldsymbol{X})_{ij} = \|\boldsymbol{X}_{ij}\|_* = \sum_{k=1}^{d} \sigma_k(\boldsymbol{X}_{ij}), \ \forall i \in [n], j \in [K].$$

We present our projection algorithm in Algorithm 2 and show that it has an $O(n \log n)$ running time in Proposition 4.

Algorithm 2 Computation of $\Pi_{\mathcal{F}}$

```
**Input:** $\boldsymbol{X} \in \mathbb{R}^{nd \times Kd}$
$\boldsymbol{H} \leftarrow \Pi_{\mathcal{H}}(\mu(\boldsymbol{X}))$
generate a sequence $\{e_i\}_{i=1}^n$ such that $\boldsymbol{H}_{ie_i} = 1$
**for** $i = 1, 2, \cdots, n$ **do**
    $\boldsymbol{R}_i \leftarrow \Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ie_i})$
**end for**
$\boldsymbol{V} \leftarrow (\mathbf{1}_{1 \times K} \otimes \boldsymbol{R}) \odot (\boldsymbol{H} \otimes \mathbf{1}_{d \times d})$
**Return:** $\boldsymbol{V}$
```

Proposition 4. *Given that $K, d = \Theta(1)$, Algorithm 2 exactly solves $\Pi_{\mathcal{F}}(\boldsymbol{X})$ in $O(n \log n)$ time.*

Let us review the design of GPM with intuition. As its name suggests, the design of GPM is inspired by the classical *power method* as well as its variant *orthogonal iteration method* (Golub and Van Loan, 2013) for obtaining the dominant eigenvector and the invariant (orthogonal) subspaces of a matrix, respectively. In fact, our design is exactly encouraged by the structural resemblances between Problem JointQP-$\mathcal{O}(d)$ and the classical eigenvalue problems above. For example, both of them aim to maximize a quadratic objective, and both of them are subject to norm and orthogonality constraints (note that $\frac{1}{\sqrt{m}}\boldsymbol{V}$ is orthonormal).

At the same time, however, the analysis of the algorithm is never a trivial corollary of the classical results because $\mathcal{F}$ has a much more complicated geometry due to its discrete structure from the community model and orthogonal block features from the synchronization model, which demands our hands-on analysis.

5 MAIN RESULTS

In this section, we formally state the main results, Theorem 1–3, on the guarantee of linear convergence of GPM under the metric of the estimation error per iteration in the logarithmic sparsity region of SGBM, *i.e.*, $p, q = O\left(\frac{\log n}{n}\right)$. We defer the proofs to the Appendix.

Definition 3. *Consider $\boldsymbol{V} \in \mathcal{F}$ corresponding to a clustering function C. For an observation matrix $\boldsymbol{A} \in \mathbb{R}^{nd \times nd}$, let $\boldsymbol{M} = \mu(\boldsymbol{AV})$. Then $\boldsymbol{A}$ is said to preserve $\boldsymbol{V}$ by δ-separation if $\boldsymbol{M}_{iC(i)} - \boldsymbol{M}_{ij} \geq \delta > 0$ for all $i \in [n]$ and $j \neq C(i), j \in [K]$.*

Remark 1. *The definition of δ-separation characterizes the signal-to-noise ratio of our observations. Specifically, for any node i, it states that the signal $\boldsymbol{M}_{iC^*(i)}$ corresponding to its true cluster $C^*(i)$ is decently separated (in terms of the nuclear norm) from the noise $\boldsymbol{M}_{ij}$ corresponding to any other cluster $j \neq C^*(i)$. This helps us bound the Lipschitz constant of the projection function $\Pi_{\mathcal{F}}(\cdot)$ in Proposition 5 in the Appendix and further establish the linear convergence of the estimation error.*

Theorem 1 (Main)**.** *Suppose that the observation matrix $\boldsymbol{A}$ is generated by*

$$\mathrm{SGBM}\left(n, p = \frac{\alpha \log n}{n}, q = \frac{\beta \log n}{n}, K, d, \mathcal{G}\right),$$

where $\alpha, \beta, K, d = \Theta(1)$*, and* $\mathcal{G} \in \{\mathcal{O}(d), \mathcal{SO}(d)\}$*. Let* $\boldsymbol{V}^0 \in \mathcal{F}$ *and* $\boldsymbol{A}$ *be the input of Algorithm 1. Let* $\boldsymbol{V}^* \in \mathcal{E} \subseteq \mathcal{F}$ *be the ground truth matrix, with* C^* *the ground truth clustering function. Then, Algorithm 1 outputs* $\boldsymbol{V}^T \in \mathcal{E}$ *such that* $\epsilon_{\mathcal{G}}(\boldsymbol{V}^T) \leq \tau$ *within* $O\left(n \log n \log \frac{n}{\tau}\right)$ *time, if*

1. *there exists a positive constant* χ *such that* $\boldsymbol{A}$ *perserves* $\boldsymbol{V}^*$ *by* (χmp)*-separation;*
2. *for all* $i \in [n]$*,* $\mu(\boldsymbol{A}\boldsymbol{V}^*)_{iC^*(i)} \geq \frac{\sqrt{2K\beta}}{\alpha} mp$*;*
3. *for* χ *satisfying (i), there exists*

$$\rho > 2\sqrt{2}\sqrt{\frac{d^2}{\chi^2} + \frac{\alpha^2}{2K\beta}}$$

such that $\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}^0) \leq \frac{\sqrt{m}}{\rho}$.

Note that Theorem 1 is a deterministic statement given the three conditions above. The first and second conditions describe a decent interaction between the observation $\boldsymbol{A}$ and the ground truth $\boldsymbol{V}^*$, which sufficiently separates the signal (mutual measurement) from noise. The third condition, on the other hand, demands a reasonable initial guess $\boldsymbol{V}^0$ of $O(\sqrt{m})$ error before the refinement steps in GPM. Since the largest possible estimation error

$$\sup_{\boldsymbol{V} \in \mathcal{F}} \epsilon_{\mathcal{O}(d)}(\boldsymbol{V}) = 2\sqrt{Kdm} = O(\sqrt{m}),$$

condition 3 is rather tolerant because it only requires constant improvement from the worst guess.

The following Theorem 2 says that conditions 1 and 2 hold with high probability, given reasonable model parameters (α, β, K, d).

Theorem 2 (Conditions 1 and 2)**.** *Suppose that* $\alpha, \beta, K, d = \Theta(1)$*, and the observation matrix* $\boldsymbol{A}$ *is generated by* SGBM *for the given ground truth matrix* $\boldsymbol{V}^*$ *with the clustering mapping* C^**. If*

$$\begin{cases} \sqrt{2K\beta} < \alpha, & (5) \\ \alpha - \sqrt{2K\beta} \log \frac{e\alpha}{\sqrt{2K\beta}} > K, & (6) \end{cases}$$

then conditions 1 and 2 in Theorem 1 happen simultaneously with probability at least $1 - n^{-\Omega(1)}$ *for a sufficiently large* n*.*

Remark 2. *The recovery requirements in* (5) *and* (6) *are converted to* δ*-separation in Theorem 2, thus implicitly controlling the linear convergence of our proposed power method.*

Condition 3 requires the existence of a good initializer. To justify this, we design the following spectral initialization method and show that it is able to generate a good $\boldsymbol{V}^0$ with high probability.[1]

Algorithm 3 Randomized spectral clustering

```
Input: the observation matrix $\boldsymbol{A}$
Initialize: $\boldsymbol{R}^0 \in \mathbb{R}^{nd \times d}$
generate $\boldsymbol{H}^0$ by Algorithm 2 in Gao et al. (2017)
$\boldsymbol{H}^0 \leftarrow \Pi_{\mathcal{H}}(\boldsymbol{H}^0)$, $\widehat{\boldsymbol{U}} \leftarrow \mathrm{eigs}_{Kd}(\boldsymbol{A})$
for $i \in [K]$ do
    pick $\tau_i \in \mathcal{I}_i^0$ uniformly randomly
    for $v \in \mathcal{I}_i^0$ do
        $\boldsymbol{R}_v^0 \leftarrow \mathrm{argmin}_{\boldsymbol{R} \in \mathcal{O}(d)} \left\| \widehat{\boldsymbol{U}}_{v\times} - \boldsymbol{R}\widehat{\boldsymbol{U}}_{\tau_i \times} \right\|_F$
    end for
end for
$\boldsymbol{V}^0 \leftarrow \left(\mathbf{1}_{1\times K} \otimes \boldsymbol{R}^0\right) \odot \left(\boldsymbol{H}^0 \otimes \mathbf{1}_{d\times d}\right)$
Return: $\boldsymbol{V}^0$
```

Theorem 3 (Condition 3)**.** *Suppose* $\alpha, \beta, K, d = \Theta(1)$*, and the observation matrix* $\boldsymbol{A}$ *is generated by* SGBM*. Then, Algorithm 3 generates an initial* $\boldsymbol{V}^0$ *satisfying condition 3 in Theorem 1 with probability at least* $1 - (\log n)^{-\Omega(1)}$ *for a sufficiently large* n*.*

6 NUMERICAL EXPERIMENTS

To corroborate the theoretical analysis, this section evaluates the performance of GPM through different numerical experiments including its phase transition behavior, convergence performance, and CPU time, and we provide a comparison with SDP (Fan et al., 2022). All the simulations are conducted via MATLAB R2021a on a workstation hosting a 64-bit Windows 10 environment with Intel(R) Xeon(R) CPU E5-2699 2.20GHz 2-processor CPU. Additional experiment results are deferred to Appendix B.

6.1 Phase Transition

We first report the phase transition behavior of GPM. For each selected pair of parameters (n, K), we increment α and β from 0 to $\frac{n}{\log n}$, and generate the observation under SGBM for $N = 50$ times. Each time, GPM is invoked to attempt to recover the ground truth. We regard an attempt *successful* if and only if $\epsilon_{\mathcal{G}}(\boldsymbol{V}^\top) \leq \tau = 10^{-3}$, and the

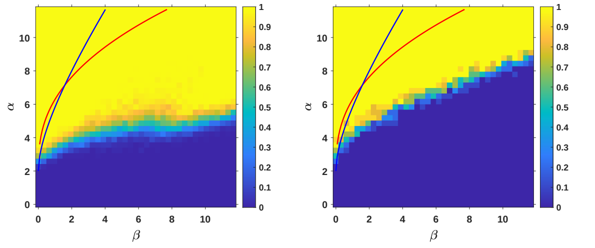
[1] Concurrent with our work, there emerged other spectral methods such as Fan et al. (2023) that are proved to produce a constant-level error for $K = 2$ case. We remark that any initializer satisfying the $\sqrt{m}/\rho$ requirement is acceptable, because GPM is always able to refine the initialization to arbitrary precision. In this paper we shall stick to our algorithm as its theory covers all $K = \Theta(1)$.

rate of success at (α, β) is defined to be $r(\alpha, \beta) = (\#\ \text{successes})/N$.

We plot $r(\alpha, \beta)$ versus the change of α and β in Figure 2, together with the theoretical threshold for pure community detection $\sqrt{\alpha} - \sqrt{\beta} = \sqrt{K}$ (blue) and the lower bound claimed in Theorem 2 (red). See the Appendix for phase transition plots under more parameter settings.

The results clearly exhibit a behavior of phase transition, and the theoretical lower bound is a sufficient control. The gap in between indicates that even the improved lower bound is still not tight for GPM, which invites further analysis of the algorithm. One can also observe that the transition pattern behaves slightly differently for small and large settings of α and β. This may suggest different properties of GPM in the logarithmic sparsity region and the linear region.

We also plot the phase transition pattern of SDP proposed in Fan et al. (2022), which is implemented by MATLAB CVX package and MOSEK solver. According to Figure 2, the recovery performance of GPM notably outperforms SDP, while the boundary of transition is less sharp.



(a) GPM

(b) SDP

Figure 2: Phase transition results on GPM and SDP with $\mathcal{G} = \mathcal{SO}(3)$ and $(n, K) = (50, 2)$. The theoretical threshold for pure community detection $\sqrt{\alpha} - \sqrt{\beta} = \sqrt{K}$ is plotted in blue; the improved lower bound claimed in Theorem 2 is plotted in red.

6.2 Convergence Performance and CPU Time

The convergence performance of GPM is also studied. At $\mathcal{G} = \mathcal{O}(d), n = 400$ and three different settings of (K, α, β), we keep track of the estimation error dynamics $\epsilon_{\mathcal{G}}(\boldsymbol{V}^t)$ of GPM. Experiments are repeated 10 times for each group of parameters, and we plot the dynamics for one parameter setting in Figure 3 and leave others in the Appendix. Typically, GPM is able to recover the ground truth within a considerably small number of iterations after a (log-scaled) linear decay of estimation error, which again aligns with our theory.

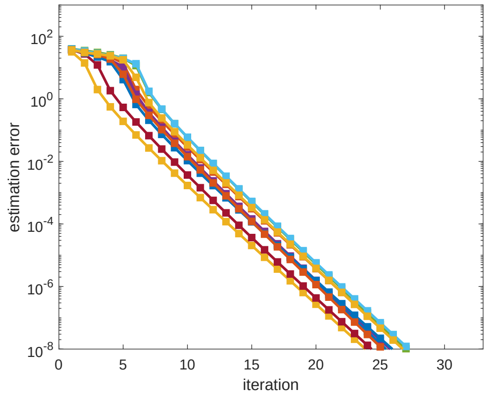


Figure 3: Convergence results of GPM at $\mathcal{G} = \mathcal{O}(d)$ and $n = 400$, with parameters $(K, \alpha, \beta) = (5, 15, 10)$. We plot 10 independent trails.

Table 1: Comparison of average CPU time (in seconds) between GPM and SDP at $\mathcal{G} = \mathcal{SO}(3), K = 2$.

parameters	CPU time (s)	
	GPM	SDP
$n = 50, \alpha = 8, \beta = 5$	**7.37**	38.69
$n = 100, \alpha = 15, \beta = 10$	**7.81**	721.21
$n = 200, \alpha = 25, \beta = 15$	**11.40**	3600+
$n = 400, \alpha = 35, \beta = 25$	**20.39**	3600+

To further evidence the superiority of GPM in respect of its time complexity, we also test the average CPU time of both GPM and SDP on problems of different scales. All the results reported in Table 1 obviously demonstrate a significantly improved time efficiency of our algorithm than SDP.

7 CONCLUSION

This paper proposes a GPM to directly tackle the non-convex problem of joint community detection and group synchronization in the logarithmic sparsity region of SGBM. From the theoretical side, we give a probabilistic bound for GPM to exactly recover the ground truth in $O(n \log^2 n)$ time, as an improvement of the previous method relying on SDP (Fan et al., 2022) that typically takes $O(n^{3.5})$ time. We also design a randomized spectral clustering method as an initializer of GPM.

Generalizing the previous SDP theory, our analysis of GPM applies to both orthogonal and rotational synchronization, and makes up the unresolved cases of $K \geq 3$. Meanwhile, our probabilistic bound breaches the information-theoretic limit for pure community detection under SBM, which implies that GPM finely exploits the additional information of group structures embedded in the joint problem.

Subsequent to the completion of our work, there arose a continuing line of studies on a variety of efficient algorithms for this problem with additional restrictions. For example, Fan et al. (2023) proposes a new spectral method with a blockwise column pivoted QR factorization to achieve exact recovery of community structure and constant-level error on the group estimation for $K = 2$. Wang and Zhao (2023) proposes a multi-frequency GPM for the phase synchronization case $\mathcal{G} = \mathcal{U}(1) \cong \mathcal{SO}(2)$ by exploiting its MLE formulation. We note that our recovery guarantee for the joint problem in Theorem 2 breaks the information-theoretic lower bound $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{K}$ for pure community detection Abbe and Sandon (2015). A future direction of theoretical interest is to explore the necessary conditions for exact recovery in the joint problem. It is also of interest to apply our methodology to regimes other than the logarithmic sparsity regime of SGBM. Overall, our work opens up further questions regarding a deeper insight into the joint problem, more efficient algorithms, and the information-theoretic lower bound.

Acknowledgements

Sijin Chen and Xiwei Cheng acknowledge the support from the 2021 Summer Research Program of the Institute of Theoretical Computer Science and Communications, CUHK. Anthony Man-Cho So is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) project CUHK 14205421.

References

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86.

Abbe, E. and Sandon, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688.

Amini, A. A. and Levina, E. (2018). On semidefinite relaxations for the block model. *Annals of Statistics*, 46(1):149–179.

Arie-Nachimson, M., Kovalsky, S. Z., Kemelmacher-Shlizerman, I., Singer, A., and Basri, R. (2012). Global motion estimation from point matches. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 81–88. IEEE.

Bajaj, C., Gao, T., He, Z., Huang, Q., and Liang, Z. (2018). SMAC: Simultaneous mapping and clustering using spectral decompositions. In *Proceedings of the 35th International Conference on Machine Learning*, pages 324–333.

Bandeira, A. S., Boumal, N., and Voroninski, V. (2016). On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Proceedings of the 29th Annual Conference on Learning Theory*, pages 361–382.

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377.

Boumal, N. (2016). Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377.

Fan, Y., Khoo, Y., and Zhao, Z. (2022). Joint community detection and rotational synchronization via semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 4(3):1052–1081.

Fan, Y., Khoo, Y., and Zhao, Z. (2023). A spectral method for joint community detection and orthogonal group synchronization. *SIAM Journal on Matrix Analysis and Applications*, 44(2):781–821.

Fan, Y. and Zhao, Z. (2019). Multi-frequency vector diffusion maps. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1843–1852.

Frank, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies:Visualization of Biological Molecules in Their Native State: Visualization of Biological Molecules in Their Native State*. Oxford University Press, USA.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(60):1–45.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*. Johns Hopkins University Press, 4th edition.

Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes Problems*, volume 30 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, UK.

Hajek, B., Wu, Y., and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming.

Ling, S. (2022). Near-optimal performance bounds for orthogonal and permutation group synchronization via spectral methods. *Applied and Computational Harmonic Analysis*, 60:20–52.

Liu, H., Yue, M.-C., and So, A. M.-C. (2017). On the estimation performance and convergence rate of the generalized power method for phase synchronization. *SIAM Journal on Optimization*, 27(4):2426–2446.

Liu, H., Yue, M.-C., and So, A. M.-C. (2023). A unified approach to synchronization problems over subgroups of the orthogonal group. *Applied and Computational Harmonic Analysis*, 66:320–372.

Singer, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36.

Singer, A., Zhao, Z., Shkolnisky, Y., and Hadani, R. (2011). Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM Journal on Imaging Sciences*, 4(2):723–759.

Stewart, G. W. (1990). Perturbation theory for the singular value decomposition. In *In SVD and Signal Processing, II: Algorithms, Analysis and Applications*, pages 99–109. Elsevier.

Tokuyama, T. and Nakano, J. (1995). Geometric algorithms for the minimum cost assignment problem. *Random Structures and Algorithms*, 6(4):393–406.

Tropp, J. A. (2011). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Wang, L. and Zhao, Z. (2023). Multi-frequency joint community detection and phase synchronization. *IEEE Transactions on Signal and Information Processing over Networks*, 9:162–174.

Wang, P., Liu, H., Zhou, Z., and So, A. M.-C. (2021). Optimal non-convex exact recovery in stochastic block model via projected power method. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10828–10838.

Wang, P., Zhou, Z., and So, A. M.-C. (2022a). Non-convex exact community recovery in stochastic block model. *Mathematical Programming*, 195(1-2):793–829.

Wang, X., Wang, P., and So, A. M.-C. (2022b). Exact community recovery over signed graphs. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 9686–9710.

Weng, H. and Feng, Y. (2016). Community detection with nodal information.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Yun, S.-Y. and Proutiere, A. (2014). Accurate community detection in the stochastic block model via spectral algorithms.

Zhang, Y., Levina, E., and Zhu, J. (2016). Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153 – 3178.

Zhao, Z. and Singer, A. (2014). Rotationally invariant image representation for viewing direction classification in cryo-em. *Journal of Structural Biology*, 186(1):153–166.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Not Applicable

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Non-Convex Joint Community Detection and Group Synchronization via Generalized Power Method: Supplementary Materials

A PROOFS

A.1 Proof of Proposition 4

Proposition 4 guarantees the correctness and efficiency of Algorithm 2. The proof relies on the invariance of the inner product $\langle \boldsymbol{X}, \boldsymbol{V} \rangle$ for any $\boldsymbol{X} \in \mathbb{R}^{nd\times Kd}$ and $\boldsymbol{V} \in \mathcal{F}$, when transitioning the *mask* $\boldsymbol{H}$ from $\boldsymbol{V}$ to $\boldsymbol{X}$:

Lemma 1. *For any $\boldsymbol{X} \in \mathbb{R}^{nd\times Kd}$, $\boldsymbol{R} \in \mathcal{O}(d)^n$, and $\boldsymbol{H} \in \mathcal{H}$,*

$$\operatorname{tr}\Big(\boldsymbol{X}^\top\big((\mathbf{1}_{1\times K} \otimes \boldsymbol{R}) \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d})\big)\Big) = \operatorname{tr}\Big(\big(\boldsymbol{X} \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d})\big)^\top (\mathbf{1}_{1\times K} \otimes \boldsymbol{R})\Big).$$

The proof of Lemma 1 is deferred to Appendix A.5

Proof of Proposition 4. We treat $\boldsymbol{X}$ as an n by K block matrix. By definition,

$$\Pi_{\mathcal{F}}(\boldsymbol{X}) = \underset{\boldsymbol{V}\in\mathcal{F}}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{V}\|_F = \underset{\boldsymbol{V}\in\mathcal{F}}{\operatorname{argmin}} \|\boldsymbol{X}\|_F^2 + \|\boldsymbol{V}\|_F^2 - \langle \boldsymbol{X}, \boldsymbol{V} \rangle,$$

where $\|\boldsymbol{V}\|_F^2 = n$ is constant for all $\boldsymbol{V} \in \mathcal{F}$. Hence,

$$\Pi_{\mathcal{F}}(\boldsymbol{X}) = \underset{\boldsymbol{V}\in\mathcal{F}}{\operatorname{argmax}} \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{V}) = \underset{\boldsymbol{R}\in\mathcal{O}(d)^n, \boldsymbol{H}\in\mathcal{H}}{\operatorname{argmax}} \operatorname{tr}\Big(\boldsymbol{X}^\top\big((\mathbf{1}_{1\times K} \otimes \boldsymbol{R}) \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d})\big)\Big).$$

Applying Lemma 1,

$$\begin{aligned}\Pi_{\mathcal{F}}(\boldsymbol{X}) &= \underset{\boldsymbol{R}\in\mathcal{O}(d)^n, \boldsymbol{H}\in\mathcal{H}}{\operatorname{argmax}} \operatorname{tr}\Big(\big(\boldsymbol{X} \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d})\big)^\top (\mathbf{1}_{1\times K} \otimes \boldsymbol{R})\big)\Big) \\ &= \underset{\boldsymbol{R}\in\mathcal{O}(d)^n, \boldsymbol{H}\in\mathcal{H}}{\operatorname{argmax}} \operatorname{tr}\left(\sum_{i=1}^n (\boldsymbol{X}_{ie_i})^\top \boldsymbol{R}_i\right), \qquad (7)\end{aligned}$$

where e_i is such that $\boldsymbol{H}_{ie_i} = 1$. For any fixed $\boldsymbol{H}$ and every $i \in [n]$, we denote $\boldsymbol{U\Sigma V}^\top$ the SVD of $\boldsymbol{X}_{ie_i}$. By Proposition 2,

$$\underset{\boldsymbol{R}_i\in\mathcal{O}(d)}{\operatorname{argmax}} \operatorname{tr}\left((\boldsymbol{X}_{ie_i})^\top \boldsymbol{R}_i\right) = \boldsymbol{UV}^\top$$

and accordingly

$$\max_{\boldsymbol{R}_i\in\mathcal{O}(d)} \operatorname{tr}\left((\boldsymbol{X}_{ie_i})^\top \boldsymbol{R}_i\right) = \operatorname{tr}(\boldsymbol{\Sigma}) = \mu(\boldsymbol{X})_{ie_i}.$$

Therefore, in order to maximize the expression in (7), it suffices to find

$$\underset{\boldsymbol{H}\in\mathcal{H}}{\operatorname{argmax}} \sum_{i=1}^n \mu(\boldsymbol{X})_{ie_i} = \underset{\boldsymbol{H}\in\mathcal{H}}{\operatorname{argmax}} \sum_{i=1}^n \sum_{j=1}^K \mu(\boldsymbol{X})_{ij} \boldsymbol{H}_{ij} = \underset{\boldsymbol{H}\in\mathcal{H}}{\operatorname{argmax}} \langle \boldsymbol{H}, \mu(\boldsymbol{X}) \rangle = \Pi_{\mathcal{H}}(\mu(\boldsymbol{X})),$$

and then to perform $\mathcal{O}(d)$ projections for the selected blocks $\boldsymbol{X}_{ie_i}$. This validates algorithm 2. Given that $K, d = \Theta(1)$, step 2 in the algorithm takes $O(n \log n)$ time according to Proposition 1, and all other steps take $O(n)$ time. This completes the proof. □

A.2 Proof of Theorem 1

To prove Theorem 1, we first show the linear convergence of Algorithm 1 in the quotient space $\mathcal{F}/\sim$ under the relaxed metric of estimation error, $\epsilon_{\mathcal{O}(d)}$. Then, we present the tightness of the rounding procedure $\mathcal{R}$ that brings forth Theorem 1 as a direct consequence.

Linear convergence under the metric $\epsilon_{\mathcal{O}(d)}$

This section is devoted to proving the following theorem on the linear convergence of Algorithm 1 under the relaxed metric of estimation error $\epsilon_{\mathcal{O}(d)}$.

Theorem 4. *Suppose that the observation matrix $\boldsymbol{A}$ is generated by*

$$\mathrm{SGBM}\left(n, \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, K, d, \mathcal{G}\right),$$

where $\alpha, \beta, K, d = \Theta(1)$, and $\mathcal{G} \in \{\mathcal{O}(d), \mathcal{SO}(d)\}$. Let $\boldsymbol{V}^0 \in \mathcal{F}$ and $\boldsymbol{A}$ be the input of Algorithm 1, and $\{\boldsymbol{V}^1, \boldsymbol{V}^2, ...\}$ a sequence generated by the iterations in Algorithm 1. Let $\boldsymbol{V}^* \in \mathcal{E} \subset \mathcal{F}$ be the ground truth matrix that incorporates the clustering function C^*. Then,

$$\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}^{t+1}) \leq \frac{1}{2}\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}^t)$$

for arbitrary non-negative integer t, if conditions (i)–(iii) in Theorem 1 hold.

Recall that GPM iteratively updates the variable by the map $\Pi_{\mathcal{F}}(\boldsymbol{A}\,\cdot\,) : \mathcal{F} \to \mathcal{F}$. Therefore, in order to prove Theorem 4, it should be crucial to identify the important properties of the projection operator $\Pi_{\mathcal{F}}$. We first make two useful observations in the following Lemmas (whose proofs are deferred to Sections A.6 and A.7):

Lemma 2. *For any $\boldsymbol{Q} \in \mathcal{P}_K(\mathcal{O}(d))$ and $\boldsymbol{X} \in \mathbb{R}^{nd \times Kd}$, $\Pi_{\mathcal{F}}(\boldsymbol{X}\boldsymbol{Q}) = \Pi_{\mathcal{F}}(\boldsymbol{X})\boldsymbol{Q}$.*

Lemma 3. *If condition (i) in Theorem 1 holds, then $\Pi_{\mathcal{F}}(\boldsymbol{A}\boldsymbol{V}^*) = \boldsymbol{V}^*$.*

The core of the proof lies in controlling the behavior of the projection operator $\Pi_{\mathcal{F}}$ so that the estimation error after each update can be bounded. One possible way, as Proposition 5 follows, is to show $\Pi_{\mathcal{F}}$ possesses a Lipschitz-like property of linearly controlling the Frobenius distance of two points after a projection. The proof of Proposition 5 is presented in Section A.8.

Proposition 5. *Let $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{V}^*$. If conditions (i) and (ii) in Theorem 1 hold, then*

$$\|\Pi_{\mathcal{F}}(\boldsymbol{X}) - \Pi_{\mathcal{F}}(\boldsymbol{X}')\|_F \leq \frac{2}{mp}\sqrt{\frac{d^2}{\chi^2} + \frac{\alpha^2}{2K\beta}}\, \|\boldsymbol{X} - \boldsymbol{X}'\|_F\,.$$

for any $\boldsymbol{X}' \in \mathbb{R}^{nd \times Kd}$.

For the sequence $\{\boldsymbol{V}^0, \boldsymbol{V}^1, \boldsymbol{V}^2, ...\}$ generated by GPM, denote $\boldsymbol{Q}^t = \mathrm{argmin}_{\boldsymbol{Q} \in \mathcal{P}_K(\mathcal{G})} \|\boldsymbol{V}^t - \boldsymbol{V}^*\boldsymbol{Q}\|_F$ for all $t \geq 0$. Given that conditions (i) and (ii) in Theorem 1 hold, we have

$$\begin{aligned}
\left\|\boldsymbol{V}^{t+1} - \boldsymbol{V}^*\boldsymbol{Q}^{t+1}\right\|_F &\leq \left\|\Pi_{\mathcal{F}}(\boldsymbol{A}\boldsymbol{V}^t) - \boldsymbol{V}^*\boldsymbol{Q}^t\right\|_F = \left\|\Pi_{\mathcal{F}}(\boldsymbol{A}\boldsymbol{V}^t\boldsymbol{Q}^{t\top}) - \boldsymbol{V}^*\right\|_F \\
&= \left\|\Pi_{\mathcal{F}}(\boldsymbol{A}\boldsymbol{V}^t\boldsymbol{Q}^{t\top}) - \Pi_{\mathcal{F}}(\boldsymbol{A}\boldsymbol{V}^*)\right\|_F \leq \frac{2}{mp}\sqrt{\frac{d^2}{\chi^2} + \frac{\alpha^2}{2K\beta}}\left\|\boldsymbol{A}\boldsymbol{V}^t\boldsymbol{Q}^{t\top} - \boldsymbol{A}\boldsymbol{V}^*\right\|_F \\
&= \frac{2}{mp}\sqrt{\frac{d^2}{\chi^2} + \frac{\alpha^2}{2K\beta}}\left\|\boldsymbol{A}(\boldsymbol{V}^t - \boldsymbol{V}^*\boldsymbol{Q}^t)\right\|_F,
\end{aligned}$$

where Lemma 2 yields the first equality, Lemma 3 yields the second equality, and the second inequality is due to Proposition 5. We can proceed to obtain

$$\begin{aligned}\left\|\boldsymbol{A}(\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t)\right\|_F &\leq \left\|(\boldsymbol{A}-p\boldsymbol{V}^*\boldsymbol{V}^{*\top})(\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t)\right\|_F+\left\|p\boldsymbol{V}^*\boldsymbol{V}^{*\top}(\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t)\right\|_F\\ &\leq \left\|\boldsymbol{A}-p\boldsymbol{V}^*\boldsymbol{V}^{*\top}\right\|\left\|\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t\right\|_F+\left\|p\boldsymbol{V}^*\boldsymbol{V}^{*\top}(\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t)\right\|_F\\ &= \left\|\boldsymbol{A}-p\boldsymbol{V}^*\boldsymbol{V}^{*\top}\right\|\left\|\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t\right\|_F+m\sqrt{mp}\left\|\frac{1}{m}(\boldsymbol{V}^*\boldsymbol{Q}^t)^\top\boldsymbol{V}^t-\boldsymbol{I}_{Kd}\right\|_F.\end{aligned}\tag{8}$$

Since we are expecting some relationship between $\left\|\boldsymbol{V}^{t+1}-\boldsymbol{V}^*\boldsymbol{Q}^{t+1}\right\|_F$ and $\|\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t\|_F$, it remains to bound the two components $\left\|\boldsymbol{A}-p\boldsymbol{V}^*\boldsymbol{V}^{*\top}\right\|$ and $\left\|\frac{1}{m}\left(\boldsymbol{V}^*\boldsymbol{Q}^t\right)^\top\boldsymbol{V}^t-\boldsymbol{I}_{Kd}\right\|_F$ respectively, as is presented in the following two Propositions (whose proofs are deferred to Sections A.9 and A.10).

Proposition 6. *There exists $c_1, c_2, c_3 > 0$ such that*

$$\left\|\boldsymbol{A}-p\boldsymbol{V}^*\boldsymbol{V}^{*\top}\right\| \leq c_1\sqrt{qm}+c_2\sqrt{pm}+c_3\sqrt{\log n}$$

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

Remark 3. *In the logarithmic sparsity region, Proposition 6 gives a $O(\sqrt{\log n})$ bound. Different from the techniques in Theorem 6 of Liu et al. (2023) for controlling a similar term, the result here does not rely on the celebrated matrix Bernstein inequality (Tropp, 2011), which loosely controls the term at $O(\log n)$.*

Proposition 7. *For $\rho > 0$, if $\|\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t\|_F \leq \frac{\sqrt{m}}{\rho}$, then*

$$m\left\|\frac{1}{m}\left(\boldsymbol{V}^*\boldsymbol{Q}^t\right)^\top\boldsymbol{V}^t-\boldsymbol{I}_{Kd}\right\|_F \leq \frac{1}{2}\sqrt{1+\frac{1}{d}}\frac{\sqrt{m}}{\rho}\left\|\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t\right\|_F.$$

Proof for Theorem 4. Invoking Proposition 6 and 7, (8) becomes

$$\begin{aligned}\left\|\boldsymbol{V}^{t+1}-\boldsymbol{V}^*\boldsymbol{Q}^{t+1}\right\|_F &\leq \frac{2}{mp}\sqrt{\frac{d^2}{\chi^2}+\frac{\alpha^2}{2K\beta}}\left(\frac{1}{2}\sqrt{1+\frac{1}{d}}\frac{mp}{\rho}+c_0\sqrt{\log n}\right)\left\|\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t\right\|_F\\ &\leq \frac{\sqrt{2}+o(1)}{\rho}\sqrt{\frac{d^2}{\chi^2}+\frac{\alpha^2}{2K\beta}}\left\|\boldsymbol{V}^t-\boldsymbol{V}^*\boldsymbol{Q}^t\right\|_F\end{aligned}$$

for a sufficiently large n. When

$$\rho > 2\sqrt{2}\sqrt{\frac{d^2}{\chi^2}+\frac{\alpha^2}{2K\beta}},$$

we have $\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}^{t+1}) \leq \frac{1}{2}\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}^t)$. When condition (iii) in Theorem 1 holds, the linear decay of $\epsilon(\boldsymbol{V}^t)$ inductively applies to arbitrary nonnegative integer t, which establishes Theorem 4. □

Tightness of rounding

It is an immediate implication of Theorem 4 that, prior to the rounding procedure, Algorithm 1 obtains $\boldsymbol{V}^T$ such that $\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}^T) \leq \tau$ within $T = O\left(\log\frac{n}{\tau}\right)$ iterations. In each iteration of GPM, the projection takes $O(n\log n)$ time according to Proposition 4, and the time complexity of matrix multiplication can also achieve $O(n\log n)$ due to their sparse structures. Therefore, the total time complexity to obtain such a $\boldsymbol{V}^T$ is $O\left(n\log n\log\frac{n}{\tau}\right)$. Compared with Theorem 1, the very difference lies in the metric of estimation error $\epsilon_{\mathcal{G}}$ after the rounding procedure $\mathcal{R}$. We handle the tightness of $\mathcal{R}$, hence the tightness of relaxation from $\mathcal{E}$ to $\mathcal{F}$ unresolved in Section 3, by the following Proposition (the proof is presented in Section A.11).

Proposition 8. *Suppose that $\mathcal{G} = \mathcal{SO}(d)$, and $\boldsymbol{V}^* \in \mathcal{E}$ is the ground truth. Then, for any $\boldsymbol{V} \in \mathcal{F}$ such that $\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}) < \sqrt{2}$, $\epsilon_{\mathcal{G}}(\mathcal{R}(\boldsymbol{V})) = \epsilon_{\mathcal{O}(d)}(\boldsymbol{V})$.*

In theory and practice, the tolerance constant τ is usually far below $\sqrt{2}$. Therefore, combined with Theorem 4, Proposition 8 completes the proof of Theorem 1.

A.3 Proof of Theorem 2

Firstly, we state two tail bounds for Bernoulli random variables and the random sum of uniformly distributed orthogonal matrices, respectively.

Lemma 4 (Hajek et al. (2016), Lemma 2)**.** *Let* $X \sim \mathrm{Binom}(m, \alpha \log n/n)$ *for* $m \in \mathbb{N}, \alpha = O(1)$*, where* $m = \frac{n}{K}$ *for some* $K > 0$*. Let* $\tau \in (0, \alpha]$*. Then for a sufficiently large* n*,*

$$\Pr\left(X \leq \frac{\tau}{K}\log n\right) = n^{-\frac{1}{K}\left(\alpha - \tau\log\left(\frac{e\alpha}{\tau}\right)+o(1)\right)}.$$

Lemma 5 (Fan et al. (2022), Theorem A.3)**.** *Suppose that* $\{u_i\}_{i=1}^m$ *and* $\{\boldsymbol{R}_i\}_{i=1}^m$ *are two finite random sequences independently and identically sampled from two independent distributions* $\mathrm{Bern}(q)$ *and* $\mathrm{Unif}(\mathcal{O}(d))$*, respectively. Let* $\boldsymbol{S} = \sum_{i=1}^m u_i \boldsymbol{R}_i$*. Then, with probability at least* $1 - n^{-c}$*,*

$$\|\boldsymbol{S}\| \leq \sqrt{2qm(c\log n + \log 2d)}\left(\sqrt{1 + \frac{c\log n + \log 2d}{18qm}} + \sqrt{\frac{c\log n + \log 2d}{18qm}}\right).$$

Remark 4. *Taking* $n = Km$ *and* $q = \beta\log n/n$ *into Lemma 5, one can show that*

$$\|\boldsymbol{S}\| \leq \sqrt{\frac{2c\beta}{K}}\left(\sqrt{1 + \frac{cK}{18\beta}} + \sqrt{\frac{cK}{18\beta}}\right)\log n \tag{9}$$

with probability at least $1 - n^{-c}$*. Considering* $18\beta \gg cK$*, (9) is simplified to* $\|\boldsymbol{S}\| \leq \sqrt{\frac{2c\beta}{K}}\log n$*. This simplification is always conducted throughout the following contents.*

Now we present a straightforward result on the model parameters.

Lemma 6. *Suppose that the positive constants* α, β, K, d *are given. let* $f(\tau) = \alpha - \tau\log\frac{e\alpha}{\tau}$ *defined on* $(0, \alpha]$*. If*

$$\left\{\begin{array}{ll} \sqrt{2K\beta} < \alpha, & (10)\\ \alpha - \sqrt{2K\beta}\log\frac{e\alpha}{\sqrt{2K\beta}} > K, & (11)\end{array}\right.$$

then there exists $\tilde{\tau} < \alpha$*,* $\tilde{c} > 1$*, and* $\chi > 0$*, such that*

$$\left\{\begin{array}{ll} \tilde{\tau} \geq \sqrt{2\tilde{c}K\beta} + \frac{\chi\alpha}{d}; & (12)\\ \alpha - \tilde{\tau}\log\frac{e\alpha}{\tilde{\tau}} > K. & (13)\end{array}\right.$$

Proof. One can observe that $f(\tau) = \alpha - \tau\log\left(\frac{e\alpha}{\tau}\right)$ monotonically decreases in $(0, \alpha]$. Therefore, the root τ^* such that $f(\tau^*) = K$ is uniquely determined in $(0, \alpha)$, and $\tau < \tau^*$ if $f(\tau) > K$ and $\tau \in (0, \alpha]$. By (10), there exists $c_1 > 1$ such that $\sqrt{2cK\beta} < \alpha$ for any $1 < c \leq c_1$. By (11), there exists $c_2 > 1$ such that for $1 < c \leq c_2$,

$$\alpha - \sqrt{2cK\beta}\log\frac{e\alpha}{\sqrt{2cK\beta}} > K,$$

and hence $\sqrt{2cK\beta} < \tau^*$ when $\sqrt{2cK\beta} < \alpha$. Pick $\tilde{c} \in (1, \min\{c_1, c_2\}]$, and $\tilde{\tau} \in (\sqrt{2\tilde{c}K\beta}, \tau^*)$. (12) and (13) immediately follow by taking $\chi = d(\tilde{\tau} - \sqrt{2\tilde{c}K\beta})/\alpha$. □

Proof of Theorem 2. Denote $\boldsymbol{M} = \mu(\boldsymbol{A}\boldsymbol{V}^*)$. We first consider the probability of two subevents defined as follows for fixed i, j, j' such that $C(i) = j \neq j'$, and then apply union bound.

$$\left\{\begin{array}{ll} \text{there exists a constant } \chi > 0 \text{ such that } \boldsymbol{M}_{ij} - \boldsymbol{M}_{ij'} \geq \chi mp; & (14)\\ \boldsymbol{M}_{ij} > \frac{\sqrt{2K\beta}}{\alpha}mp. & (15)\end{array}\right.$$

Observe that

$$[\boldsymbol{A}\boldsymbol{V}^*]_{ij} = \sum_{k:C(k)=j}\boldsymbol{A}_{ik}\boldsymbol{R}_k^* = \sum_{\substack{k:C(k)=j\\C(k)=C(i)}} w_{ik}\boldsymbol{R}_i^* + \sum_{\substack{k:C(k)=j\\C(k)\neq C(i)}} u_{ik}\boldsymbol{R}_{ik}\boldsymbol{R}_k^*,$$

where $w_{ik} \sim \mathrm{Bern}(p), u_{ik} \sim \mathrm{Bern}(q), \boldsymbol{R}_{ik} \sim \mathrm{Unif}(\mathcal{O}(d))$. In fact, the two parts in the summation are complementary, i.e.

$$[\boldsymbol{A}\boldsymbol{V}^*]_{ij} = \begin{cases} \left(\sum_{k:C(k)=j} w_{ik}\right) \boldsymbol{R}_i^*, & \text{if } C(i)=j, \\ \sum_{k:C(k)=j} u_{ik}\boldsymbol{R}_{ik}\boldsymbol{R}_k^*, & \text{otherwise.} \end{cases} \tag{16}$$

Therefore, due to edge independence, $\boldsymbol{M}_{ij} = \|X\boldsymbol{R}_i^*\|_* = dX$ where $X \sim \mathrm{Binom}(m, \alpha \log n/n)$. Likewise, denoting $\boldsymbol{S}$ the random variable as stated in Lemma 5, $\sigma_1(\boldsymbol{X}_{ij'}) = \|\boldsymbol{S}\|$ since the distribution of $\mathrm{Unif}(\mathcal{O}(d))$ is invariant under right (and left) orthogonal group actions, and consequently $\boldsymbol{M}_{ij'} \le d\sigma_1(\boldsymbol{X}_{ij'}) = d\,\|\boldsymbol{S}\|$. Then both (14) and (15) are guaranteed to happen when

$$\begin{cases} X - \|\boldsymbol{S}\| \ge \frac{\chi}{d}mp = \frac{\chi\alpha}{Kd}\log n; & (17) \\ \|X\| \ge \sqrt{\frac{2\beta}{K}}\log n. & (18) \end{cases}$$

With (5) and (6), we are able to invoke Lemma 6 to find a group of parameters $\tilde{\tau}, \tilde{c}, \chi$ such that

$$\begin{cases} \tilde{\tau} < \alpha, \tilde{c} > 1, \chi > 0; & (19) \\ \tilde{\tau} \ge \sqrt{2\tilde{c}K\beta} + \frac{\chi\alpha}{d}; & (20) \\ \alpha - \tilde{\tau}\log\frac{e\alpha}{\tilde{\tau}} > K. & (21) \end{cases}$$

Then, Lemma 4 indicates that

$$\Pr\left(X \ge \frac{\tilde{\tau}}{K}\log n\right) \ge 1 - n^{-\frac{1}{K}\left(\alpha - \tilde{\tau}\log\frac{e\alpha}{\tilde{\tau}}\right)}, \tag{22}$$

while another probabilistic bound on $\|\boldsymbol{S}\|$ is derived from Lemma 5:

$$\Pr\left(\|\boldsymbol{S}\| \le \sqrt{\frac{2\tilde{c}\beta}{K}\log n}\right) \ge 1 - n^{-\tilde{c}}. \tag{23}$$

Combined with (19) and (20), the two events in (22) and (23) would immediately imply (17) and (18), and consequently the subevents (14) and (15). They would further establish the final proposition, given that the probability of both events stated in (22) and (23) is sufficiently high even after taking union bound over all $i \in [n]$ and $j' \in [K]$. This is guaranteed by (19) and (21) because, by union bound, both events hold for all i, j' with probability at least

$$1 - nKn^{-\frac{1}{K}\left(\alpha - \tilde{\tau}\log\frac{e\alpha}{\tilde{\tau}}\right)} - n^{-\tilde{c}+1} = 1 - Kn^{-\frac{1}{K}\left(\alpha - \tilde{\tau}\log\frac{e\alpha}{\tilde{\tau}} - K\right)} - n^{-\tilde{c}+1} = 1 - n^{-\Omega(1)}.$$

This completes the proof. □

A.4 Proof of Theorem 3

While the community structure and group information are jointly optimized in GPM for exact recovery, this initialization algorithm adopts an intuitive two-stage design as condition (*iii*) in Theorem 1 tolerates a rough estimation. The first stage generates a preliminary guess of the community structure $\boldsymbol{H}$. Similar to Wang et al. (2021), we invoke Algorithm 2 in Gao et al. (2017), a greedy spectral clustering method, to obtain a (imbalanced) clustering matrix, which is then rounded to a balanced clustering matrix $\boldsymbol{H}^0 \in \mathcal{H}$. Theory developed in Wang et al. (2021) has shown that for any numerical constant C,

$$\epsilon(\boldsymbol{H}^0) \le \sqrt{\frac{Cn}{\log n}} \tag{24}$$

with probability at least $1 - n^{-\Omega(1)}$. After obtaining an initialization $\boldsymbol{H}$, the group elements are estimated in the second stage. For each hypothesized cluster i obtained previously, a pivot node denoted by τ_i is located by a randomized mechanism. Then, the relative group transformation between the pivot and any other node is estimated utilizing the corresponding block rows of $\widehat{\boldsymbol{U}} = \operatorname{eigs}_{Kd}(\boldsymbol{A})$. We consolidate the two estimations into a matrix $\boldsymbol{V}^0 \in \mathcal{F}$ as the return of the algorithm.

As the following Proposition points out, the algorithm has a decent theoretical performance regarding the estimation precision of $\boldsymbol{V}^0$ it generates. The proof of Proposition 9 is deferred to Section A.12.

Proposition 9. *Suppose that $\alpha, \beta, K, d = \Theta(1)$, and the observation matrix $\boldsymbol{A}$ is generated by* SGBM *for the given ground truth matrix $\boldsymbol{V}^*$. Then, for any given constant $\rho > 0$, Algorithm 3 generates a $\boldsymbol{V}^0$ such that*

$$\epsilon_{\mathcal{O}(d)}\left(\boldsymbol{V}^0\right) \le \frac{\sqrt{m}}{\rho}$$

with probability at least $1 - (\log n)^{-\Omega(1)}$ for a sufficiently large n.

Proof of Theorem 3. In fact, equipped with Proposition 9, Theorem 3 is immediately established. However, we remark that this is not necessarily the unique algorithm fulfilling the requirements therein. □

Remark 5. *We also note that for $K, d = \Theta(1)$, step 4 of Algorithm 3 runs in $O(n \log n)$ time by using the Lanczos method to compute the top Kd eigenvectors of a sparse observation matrix $\boldsymbol{A}$ with $O(n \log n)$ non-zero blocks. So, given an initial guess on $\boldsymbol{H}^0$, Algorithm 3 runs in $O(n \log n)$ time, which does not impose significant overhead on the power iterations in Algorithm 1 with the complexity $O(n \log^2 n)$.*

A.5 Proof of Lemma 1

Proof. We note that for any $\boldsymbol{X} \in \mathbb{R}^{nd \times Kd}$, $\boldsymbol{R} \in \mathcal{O}(d)^n$, and $\boldsymbol{H} \in \mathcal{H}$,

$$\begin{aligned}\operatorname{tr}\Big(\boldsymbol{X}^\top\big((\mathbf{1}_{1\times K} \otimes \boldsymbol{R}) \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d})\big)\Big) &= \sum_{i=1}^{n}\sum_{j=1}^{k}\langle \boldsymbol{X}_{ij}, \boldsymbol{H}_{ij}\boldsymbol{R}_i\rangle = \sum_{i=1}^{n}\sum_{j=1}^{k}\langle \boldsymbol{H}_{ij}\boldsymbol{X}_{ij}, \boldsymbol{R}_i\rangle \\ &= \operatorname{tr}\Big(\big(\boldsymbol{X} \odot (\boldsymbol{H} \otimes \mathbf{1}_{d\times d})\big)^\top (\mathbf{1}_{1\times K} \otimes \boldsymbol{R})\Big),\end{aligned}$$

where $\boldsymbol{X}_{ij}$ denotes the (i,j)-th block of $\boldsymbol{X}$ and $\boldsymbol{R}_i$ denotes the i-th block of $\boldsymbol{R}$. This completes the proof. □

A.6 Proof of Lemma 2

Proof. Since $\boldsymbol{Q} \in \mathcal{P}_K(\mathcal{O}(d))$, there exists a permutation π on $[K]$ such that $\boldsymbol{Q}_{\pi(i)i} \in \mathcal{O}(d)$ for all $i \in [K]$, and the remaining blocks of $\boldsymbol{Q}$ are zero. For any $\boldsymbol{X} = [\boldsymbol{X}_{ij}]$, we have

$$(\boldsymbol{X}\boldsymbol{Q})_{ij} = \sum_{l=1}^{K} \boldsymbol{X}_{il}\boldsymbol{Q}_{lj} = \boldsymbol{X}_{i\pi(j)}\boldsymbol{Q}_{\pi(j)j}.$$

Therefore, $\mu(\boldsymbol{X}\boldsymbol{Q})_{ij} = \left\|\boldsymbol{X}_{i\pi(j)}\right\|_* = \mu(\boldsymbol{X})_{i\pi(j)}$, and $\mu(\boldsymbol{X}\boldsymbol{Q})$ is in fact the column permutation of $\mu(\boldsymbol{X})$ according to π. Denote $\boldsymbol{H}', \boldsymbol{H}$ the clustering matrices generated in the projection algorithm on the input $\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{Q}$ respectively. Then, $\boldsymbol{H}'_{ij} = 1$ if and only if $\boldsymbol{H}_{i\pi(j)} = 1$. Now, for those i, j such that $\boldsymbol{H}'_{ij} = 1$, we have

$$\begin{aligned}\Pi_{\mathcal{F}}(\boldsymbol{X}\boldsymbol{Q})_{ij} &= \Pi_{\mathcal{O}(d)}\left((\boldsymbol{X}\boldsymbol{Q})_{ij}\right) = \Pi_{\mathcal{O}(d)}\left(\boldsymbol{X}_{i\pi(j)}\boldsymbol{Q}_{\pi(j)j}\right) \\ &= \Pi_{\mathcal{O}(d)}\left(\boldsymbol{X}_{i\pi(j)}\right)\boldsymbol{Q}_{\pi(j)j} = \Pi_{\mathcal{F}}(\boldsymbol{X})_{i\pi(j)}\boldsymbol{Q}_{\pi(j)j} = \sum_{l=1}^{K}\Pi_{\mathcal{F}}(\boldsymbol{X})_{il}\boldsymbol{Q}_{lj}.\end{aligned}$$

Hence $\Pi_{\mathcal{F}}(\boldsymbol{X}\boldsymbol{Q}) = \Pi_{\mathcal{F}}(\boldsymbol{X})\boldsymbol{Q}$. □

A.7 Proof of Lemma 3

Proof. Denote $\boldsymbol{H}'$ and $\boldsymbol{H}^*$ the clustering matrices determined in the projection algorithm on the input $\boldsymbol{A}\boldsymbol{V}^*$ and $\boldsymbol{V}^*$, respectively. Since condition (i) holds, $\boldsymbol{H}' = \boldsymbol{H}^*$. By (16), $\Pi_{\mathcal{O}(d)}\left((\boldsymbol{A}\boldsymbol{V}^*)_{iC(i)}\right) = \boldsymbol{R}_i^*$. Hence $\Pi_{\mathcal{F}}(\boldsymbol{A}\boldsymbol{V}^*) = \boldsymbol{V}^*$. □

A.8 Proof of Proposition 5

In order to establish the Lipschitz-like property of $\Pi_{\mathcal{F}}$, we first show that the maps μ and $\Pi_{\mathcal{O}(d)}$ involved in the computation of $\Pi_{\mathcal{F}}$ have a similar behavior.

Lemma 7. *For any $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{R}^{nd\times Kd}$,*

$$\|\mu(\boldsymbol{X}) - \mu(\boldsymbol{X}')\|_F \leq \sqrt{d}\,\|\boldsymbol{X} - \boldsymbol{X}'\|_F\,.$$

Proof. For simplicity we denote $\sigma_k = \sigma_k(\boldsymbol{X}_{ij})$ and $\sigma_k' = \sigma_k(\boldsymbol{X}_{ij}')$. Then

$$\begin{aligned}|\mu(\boldsymbol{X})_{ij} - \mu(\boldsymbol{X}')_{ij}| &= |\sigma_1 - \sigma_1' + \sigma_2 - \sigma_2' + ... + \sigma_d - \sigma_d'| \\ &\leq \sqrt{d}\sqrt{(\sigma_1 - \sigma_1')^2 + (\sigma_2 - \sigma_2')^2 + ... + (\sigma_d - \sigma_d')^2} \\ &\leq \sqrt{d}\left\|\boldsymbol{X}_{ij} - \boldsymbol{X}_{ij}'\right\|_F,\end{aligned}$$

where Mirsky's inequality (Stewart, 1990) yields the final step. Summing over the indices yields the desired result. □

Lemma 8. *If $\boldsymbol{X}_{ij} = \eta\boldsymbol{R}$ where $\eta > 0$ and $\boldsymbol{R} \in \mathcal{O}(d)$, then*

$$\left\|\Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}) - \Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}')\right\|_F \leq \frac{2}{\eta}\left\|\boldsymbol{X}_{ij} - \boldsymbol{X}_{ij}'\right\|_F$$

for any $\boldsymbol{X}_{ij}' \in \mathbb{R}^{d\times d}$.

Proof. Note that $\Pi_{\mathcal{O}(d)}(\boldsymbol{X}) = \Pi_{\mathcal{O}(d)}(\lambda\boldsymbol{X})$ for any $\lambda > 0$ and $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, hence we have

$$\begin{aligned}\left\|\Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}) - \Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}')\right\|_F &= \left\|\boldsymbol{R} - \Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}')\right\|_F \leq \left\|\boldsymbol{R} - \frac{1}{\eta}\boldsymbol{X}_{ij}'\right\|_F + \left\|\frac{1}{\eta}\boldsymbol{X}_{ij}' - \Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}')\right\|_F \\ &= \left\|\boldsymbol{R} - \frac{1}{\eta}\boldsymbol{X}_{ij}'\right\|_F + \left\|\frac{1}{\eta}\boldsymbol{X}_{ij}' - \Pi_{\mathcal{O}(d)}\left(\frac{1}{\eta}\boldsymbol{X}_{ij}'\right)\right\|_F \\ &\leq 2\left\|\boldsymbol{R} - \frac{1}{\eta}\boldsymbol{X}_{ij}'\right\|_F = \frac{2}{\eta}\left\|\boldsymbol{X}_{ij} - \boldsymbol{X}_{ij}'\right\|_F.\end{aligned}$$

□

Proof of Proposition 5. Let $\Pi_{\mathcal{F}}(\boldsymbol{X})$ incoporate a community structure $\boldsymbol{H}$ and $\Pi_{\mathcal{F}}(\boldsymbol{X}')$ incoporate $\boldsymbol{H}'$. Then one can observe

$$\|\Pi_{\mathcal{F}}(\boldsymbol{X}) - \Pi_{\mathcal{F}}(\boldsymbol{X}')\|_F^2 = d\,\|\boldsymbol{H} - \boldsymbol{H}'\|_F^2 + \sum_j \sum_{i\in\mathcal{I}_j\cap\mathcal{I}_j'} \left\|\Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}) - \Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}')\right\|_F^2.$$

By lemma 3 in Wang et al. (2021), lemma 7, and lemma 8,

$$\begin{aligned}\|\Pi_{\mathcal{F}}(\boldsymbol{X}) - \Pi_{\mathcal{F}}(\boldsymbol{X}')\|_F^2 &\leq \frac{4d}{\delta^2}\|\boldsymbol{M} - \boldsymbol{M}'\|_F^2 + \sum_j \sum_{i\in\mathcal{I}_j\cap\mathcal{I}_j'} \left\|\Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}) - \Pi_{\mathcal{O}(d)}(\boldsymbol{X}_{ij}')\right\|_F^2 \\ &\leq \frac{4d^2}{\delta^2}\|\boldsymbol{X} - \boldsymbol{X}'\|_F^2 + \frac{4}{\eta^2}\sum_j \sum_{i\in\mathcal{I}_j\cap\mathcal{I}_j'} \left\|\boldsymbol{X}_{ij} - \boldsymbol{X}_{ij}'\right\|_F^2 \\ &\leq \left(\frac{4d^2}{\delta^2} + \frac{4}{\eta^2}\right)\|\boldsymbol{X} - \boldsymbol{X}'\|_F^2.\end{aligned}$$

Taking $\delta = \chi mp$ and $\eta = \frac{\sqrt{2K\beta}}{\alpha}mp$ yields the result. □

A.9 Proof of Proposition 6

Proof. This is a direct generalization of Lemma 3.6 in Fan et al. (2022) when $K \geq 2$ and the constraint is relaxed from $\mathcal{SO}(d)$ to $\mathcal{O}(d)$. We apply similar notations. Observe that

$$\boldsymbol{S}_{\text{out}} = \begin{pmatrix} \mathbf{0} & \boldsymbol{S}_{12} & \boldsymbol{S}_{13} & \cdots & \boldsymbol{S}_{1K} \\ \boldsymbol{S}_{12}^\top & \mathbf{0} & \boldsymbol{S}_{23} & \cdots & \boldsymbol{S}_{2K} \\ \boldsymbol{S}_{13}^\top & \boldsymbol{S}_{23}^\top & \mathbf{0} & \cdots & \boldsymbol{S}_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_{1K}^\top & \boldsymbol{S}_{2K}^\top & \boldsymbol{S}_{3K}^\top & \cdots & \mathbf{0} \end{pmatrix} = \sum_{j>i} \begin{pmatrix} & & \\ & & \boldsymbol{S}_{ij} \\ & & \\ & \boldsymbol{S}_{ij}^\top & \\ & & \end{pmatrix}.$$

Therefore, $\|\boldsymbol{S}_{\text{out}}\| \leq \sum_{j>i} \|\boldsymbol{S}_{ij}\|$, and likewise $\|\boldsymbol{S}_{\text{in}}\| \leq \sum_i \|\boldsymbol{S}_{ii}\|$. Following the argument therein, the result can be established by union bound. □

A.10 Proof of Proposition 7

Proof. We denote for simplicity $\boldsymbol{V}^e = \boldsymbol{V}^* \boldsymbol{Q}^t$, and $\boldsymbol{Z} = \frac{1}{m} \boldsymbol{V}^{e\top} \boldsymbol{V}^t$. Recall that $\boldsymbol{V}^e$ is the optimal approximation of $\boldsymbol{V}^t$, so any per-cluster orthogonal transformation never yields a smaller difference. Specifically,

$$\left\| \boldsymbol{V}^t - \boldsymbol{V}^e \right\|_F^2 = \min_{\boldsymbol{P}} \left\| \boldsymbol{V}^t - \boldsymbol{V}^e \boldsymbol{P}^\top \right\|_F^2$$

subject to

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_1 & & & \\ & \boldsymbol{P}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{P}_K \end{pmatrix} \in \mathcal{P}_K(\mathcal{O}(d)), \boldsymbol{P}_i \in \mathcal{O}(d).$$

However, note that

$$\begin{aligned} \min_{\boldsymbol{P}} \left\| \boldsymbol{V}^t - \boldsymbol{V}^e \boldsymbol{P}^\top \right\|_F^2 &= \sum_{i=1}^K \min_{\boldsymbol{P}_i \in \mathcal{O}(d)} \left(\sum_{j \in \mathcal{I}_i^e \cap \mathcal{I}_i} \left\| \boldsymbol{V}_{ji}^t - \boldsymbol{V}_{ji}^e \boldsymbol{P}_i^\top \right\|_F^2 + \sum_{j \in \mathcal{I}_i^e \cup \mathcal{I}_i / \mathcal{I}_i^e \cap \mathcal{I}_i} \left\| \boldsymbol{V}_{ji}^t - \boldsymbol{V}_{ji}^e \boldsymbol{P}_i^\top \right\|_F^2 \right) \\ &= \sum_{i=1}^K \left(2md - 2 \max_{\boldsymbol{P}_i \in \mathcal{O}(d)} \sum_{j \in \mathcal{I}_i^e \cap \mathcal{I}_i} \left\langle \boldsymbol{V}_{ji}^t, \boldsymbol{V}_{ji}^e \boldsymbol{P}_i^\top \right\rangle \right) \\ &= \sum_{i=1}^K \left(2md - 2 \max_{\boldsymbol{P}_i \in \mathcal{O}(d)} \left\langle \boldsymbol{V}_{ji}^t, \sum_{j \in \mathcal{I}_i^e \cap \mathcal{I}_i} \boldsymbol{V}_{ji}^e \boldsymbol{P}_i^\top \right\rangle \right) \\ &= \sum_{i=1}^K \left(2md - 2m \max_{\boldsymbol{P}_i \in \mathcal{O}(d)} \left\langle \boldsymbol{P}_i, \boldsymbol{Z}_{ii}^\top \right\rangle \right), \end{aligned}$$

we obtain $\boldsymbol{P}_i = \Pi_{\mathcal{O}(d)} \left(\boldsymbol{Z}_{ii}^\top \right)$ (notice that $\boldsymbol{P}_i = \boldsymbol{I}_d$ at the same time). By Proposition 2, we have

$$\operatorname{tr}(\boldsymbol{Z}_{ii}) = \left\langle \boldsymbol{I}, \boldsymbol{Z}_{ii}^\top \right\rangle = \left\langle \Pi_{\mathcal{O}(d)} \left(\boldsymbol{Z}_{ii}^\top \right), \boldsymbol{Z}_{ii}^\top \right\rangle = \sum_{k=1}^d \sigma_k \left(\boldsymbol{Z}_{ii} \right); \tag{25}$$

$$\left\| \boldsymbol{Z}_{ii} \right\|_F^2 = \left\| \Pi_{\mathcal{O}(d)} \left(\boldsymbol{Z}_{ii}^\top \right) \boldsymbol{Z}_{ii} \right\|_F^2 = \sum_{k=1}^d \sigma_k \left(\boldsymbol{Z}_{ii} \right)^2; \tag{26}$$

$$\sigma_k(\boldsymbol{Z}_{ii}) \in [0, 1]. \tag{27}$$

We now claim that

$$\left\| \boldsymbol{Z} - \boldsymbol{I}_{Kd} \right\|_F^2 \leq \left(1 + \frac{1}{d} \right) (Kd - \operatorname{tr}(\boldsymbol{Z}))^2.$$

To this end, observe that

$$\begin{aligned}\|\boldsymbol{Z}-\boldsymbol{I}_{Kd}\|_F^2 &= \sum_{i=1}^K\left(\|\boldsymbol{Z}_{ii}\|_F^2+\sum_{j\neq i}\|\boldsymbol{Z}_{ij}\|_F^2-2\mathrm{tr}(\boldsymbol{Z}_{ii})+d\right)\\ &\leq \sum_{i=1}^K\left(\|\boldsymbol{Z}_{ii}\|_F^2+\left(\sum_{j\neq i}\|\boldsymbol{Z}_{ij}\|_F\right)^2-2\mathrm{tr}(\boldsymbol{Z}_{ii})+d\right)\\ &\leq \sum_{i=1}^K\left(\|\boldsymbol{Z}_{ii}\|_F^2+\left(\sqrt{d}-\|\boldsymbol{Z}_{ii}\|_F\right)^2-2\mathrm{tr}(\boldsymbol{Z}_{ii})+d\right). \qquad (28)\end{aligned}$$

By (25), (26), and (27), we have

$$\|\boldsymbol{Z}_{ii}\|_F^2-2\mathrm{tr}(\boldsymbol{Z}_{ii})+d=\sum_{k=1}^d\sigma_k^2-2\sum_{k=1}^d\sigma_k+d=\sum_{k=1}^d(1-\sigma_k)^2\leq\left(\sum_{k=1}^d(1-\sigma_k)\right)^2=(d-\mathrm{tr}(\boldsymbol{Z}_{ii}))^2, \qquad (29)$$

and

$$\left(\sqrt{d}-\|\boldsymbol{Z}_{ii}\|_F\right)^2\leq\left(\sqrt{d}-\frac{\mathrm{tr}(\boldsymbol{Z}_{ii})}{\sqrt{d}}\right)^2=\frac{1}{d}\left(d-\mathrm{tr}(\boldsymbol{Z}_{ii})\right)^2. \qquad (30)$$

Summing (29) and (30) over i, (28) yields

$$\begin{aligned}\|\boldsymbol{Z}-\boldsymbol{I}_{Kd}\|_F^2 &\leq \left(1+\frac{1}{d}\right)\sum_{i=1}^K(d-\mathrm{tr}(\boldsymbol{Z}_{ii}))^2\leq\left(1+\frac{1}{d}\right)\left(\sum_{i=1}^K(d-\mathrm{tr}(\boldsymbol{Z}_{ii}))\right)^2\\ &=\left(1+\frac{1}{d}\right)\left(Kd-\mathrm{tr}(\boldsymbol{Z})\right)^2,\end{aligned}$$

which validates our claim. Finally, since $\|\boldsymbol{V}^t-\boldsymbol{V}^e\|_F^2=2mKd-2m\mathrm{tr}(\boldsymbol{Z})$, we have

$$m\|\boldsymbol{Z}-\boldsymbol{I}_{Kd}\|_F\leq\frac{1}{2}\sqrt{1+\frac{1}{d}}\|\boldsymbol{V}^t-\boldsymbol{V}^e\|_F^2\leq\frac{1}{2}\sqrt{1+\frac{1}{d}}\frac{\sqrt{m}}{\rho}\|\boldsymbol{V}^t-\boldsymbol{V}^e\|_F$$

for $\rho>0$. □

A.11 Proof of Proposition 8

Proof. We first observe that the community structures of $\boldsymbol{V}$ and $\boldsymbol{V}^*$ are identical up to some permutation when $\epsilon_{\mathcal{O}(d)}(\boldsymbol{V})<\sqrt{2}$. Otherwise, at least one node falls in a erroneous cluster and $\epsilon_{\mathcal{O}(d)}(\boldsymbol{V})\geq\sqrt{2d}\geq\sqrt{2}$. Now, without loss of generality, we may identify the community structure of $\boldsymbol{V}^*$ with that of $\boldsymbol{V}$. Then no permutation is required to present the equivalence class of $\boldsymbol{V}^*$, hence

$$\epsilon_{\mathcal{O}(d)}(\boldsymbol{V})=\min_{\boldsymbol{W}\in\mathrm{bdiag}(\mathcal{O}(d)^K)}\|\boldsymbol{V}-\boldsymbol{V}^*\boldsymbol{W}\|_F, \qquad (31)$$

$$\epsilon_{\mathcal{SO}(d)}(\mathcal{R}(\boldsymbol{V}))=\min_{\boldsymbol{W}\in\mathrm{bdiag}(\mathcal{SO}(d)^K)}\|\mathcal{R}(\boldsymbol{V})-\boldsymbol{V}^*\boldsymbol{W}\|_F. \qquad (32)$$

Our second observation is that no two group elements in the same cluster of $\boldsymbol{V}$, say $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$, belong to $\mathcal{SO}(d)$ and $\mathcal{SO}^-(d)$ respectively. To see this, consider $\boldsymbol{V}^e:=\boldsymbol{V}^*\boldsymbol{W}$ where $\boldsymbol{W}\in\mathrm{bdiag}(\mathcal{O}(d)^K)$ is arbitrary. Observe that no two group elements in the same cluster of $\boldsymbol{V}^e$ belong to $\mathcal{SO}(d)$ and $\mathcal{SO}^-(d)$ respectively, because $\boldsymbol{W}$ exerts a unified group action on each cluster of $\boldsymbol{V}^*\in\mathcal{E}$. If the same does not hold for $\boldsymbol{V}$, there must exist some $i\in[n]$ such that $\boldsymbol{R}_i\in\mathcal{SO}(d),\boldsymbol{R}_i^e\in\mathcal{SO}^-(d)$, or $\boldsymbol{R}_i\in\mathcal{SO}^-(d),\boldsymbol{R}_i^e\in\mathcal{SO}(d)$. In both cases, $\|\boldsymbol{R}_i-\boldsymbol{R}_i^e\|_F\geq\sqrt{2}$, $\forall\boldsymbol{W}\in\mathrm{bdiag}(\mathcal{O}(d)^K)$, hence $\epsilon_{\mathcal{O}(d)}(\boldsymbol{V})\geq\sqrt{2}$.

Therefore, when $\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}) < \sqrt{2}$, the rounding procedure $\mathcal{R}(\boldsymbol{V}) = \boldsymbol{V}\boldsymbol{T}$, where

$$\boldsymbol{T} = \begin{pmatrix} D_1\boldsymbol{I}_d & & & \\ & D_2\boldsymbol{I}_d & & \\ & & \ddots & \\ & & & D_K\boldsymbol{I}_d \end{pmatrix} \in \mathrm{bdiag}(\mathcal{O}(d)^K),\ D_k = \det(\boldsymbol{R}_i)\ \forall i \in \mathcal{I}_k.$$

This together with (32) gives

$$\epsilon_{\mathcal{SO}(d)}(\mathcal{R}(\boldsymbol{V})) = \min_{\boldsymbol{W}\in\mathrm{bdiag}(\mathcal{SO}(d)^K)} \left\|\boldsymbol{V}\boldsymbol{T} - \boldsymbol{V}^*\boldsymbol{W}\right\|_F. \tag{33}$$

Moreover, since $\boldsymbol{T} \in \mathrm{bdiag}(\mathcal{O}(d)^n)$, (31) gives

$$\epsilon_{\mathcal{O}(d)}(\boldsymbol{V}) = \min_{\boldsymbol{W}\in\mathrm{bdiag}(\mathcal{O}(d)^K)} \left\|\boldsymbol{V} - \boldsymbol{V}^*\boldsymbol{W}\right\|_F = \min_{\boldsymbol{W}\in\mathrm{bdiag}(\mathcal{O}(d)^K)} \left\|\boldsymbol{V}\boldsymbol{T} - \boldsymbol{V}^*\boldsymbol{W}\right\|_F. \tag{34}$$

Denote $\boldsymbol{W}^*$ a minimizer of (34), *i.e.*,

$$\boldsymbol{W}^* = \operatorname*{argmin}_{\boldsymbol{W}\in\mathrm{bdiag}(\mathcal{O}(d)^K)} \left\|\boldsymbol{V}\boldsymbol{T} - \boldsymbol{V}^*\boldsymbol{W}\right\|_F.$$

Then $\left\|\boldsymbol{V}\boldsymbol{T} - \boldsymbol{V}^*\boldsymbol{W}^*\right\|_F < \sqrt{2}$. Since $\boldsymbol{V}\boldsymbol{T} \in \mathcal{E}$, it follows from an argument similar to the second observation that $\boldsymbol{W}^* \in \mathrm{bdiag}(\mathcal{SO}(d)^K)$, lest the estimation error exceeds $\sqrt{2}$. Therefore, $\boldsymbol{W}^*$ also minimizes (33). We then establish the equality $\epsilon_{\mathcal{SO}(d)}(\boldsymbol{V}\boldsymbol{T}) = \epsilon_{\mathcal{O}(d)}(\boldsymbol{V})$. □

A.12 Proof of Proposition 9

We make use of the following variant of Davis-Kahan theorem on the distance between eigenspaces of two real symmetric matrices.

Proposition 10 (Davis-Kahan theorem, Yu et al. (2015))**.** *Let $\boldsymbol{M}, \boldsymbol{M}^* \in \mathbb{R}^{N\times N}$ be symmetric matrices with eigenvalues $\lambda_1 \geq ... \geq \lambda_N$ and $\lambda_1^* \geq ... \geq \lambda_N^*$, respectively. For any integers k, l such that $1 \leq k \leq l \leq N$, let $\boldsymbol{U} = \mathrm{eigs}_{[k:l]}(\boldsymbol{M}), \boldsymbol{U}^* = \mathrm{eigs}_{[k:l]}(\boldsymbol{M}^*)$. Suppose that $\min\{\lambda_{k-1}^* - \lambda_k^*, \lambda_l^* - \lambda_{l+1}^*\} > 0$, where $\lambda_0 = +\infty$ and $\lambda_{N+1} = -\infty$. Then, there exists $\boldsymbol{Q}^* \in \mathcal{O}(l-k+1)$ such that*

$$\left\|\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{Q}^*\right\|_F \leq \frac{2\sqrt{2}\sqrt{l-k+1}\left\|\boldsymbol{M} - \boldsymbol{M}^*\right\|}{\min\{\lambda_{k-1}^* - \lambda_k^*, \lambda_l^* - \lambda_{l+1}^*\}}.$$

Proof of Proposition 9. We prove the existence of such an algorithm by showing that Algorithm 3 does satisfy all the desired properties. Observe that $\boldsymbol{V}^*/\sqrt{m}$ are the leading eigenvectors of $p\boldsymbol{V}^*\boldsymbol{V}^{*\top}$ with eigenvalues pm, while the other eigenvalues of $p\boldsymbol{V}^*\boldsymbol{V}^{*\top}$ are all zero. By Proposition 10, there exists $\boldsymbol{Q}^* \in \mathcal{O}(Kd)$ such that

$$\left\|\widehat{\boldsymbol{U}} - \frac{1}{\sqrt{m}}\boldsymbol{V}^*\boldsymbol{Q}^*\right\|_F \leq \frac{2\sqrt{2Kd}}{pm}\left\|\boldsymbol{A} - p\boldsymbol{V}^*\boldsymbol{V}^{*\top}\right\|.$$

Denote $\boldsymbol{\Phi} = \frac{1}{\sqrt{m}}\boldsymbol{V}^*\boldsymbol{Q}^*$. By Proposition 6, for sufficiently large n, there exists $c_4 > 0$ such that

$$\left\|\widehat{\boldsymbol{U}} - \boldsymbol{\Phi}\right\|_F \leq \frac{2\sqrt{2Kd}}{pm}\left(c_1\sqrt{qm} + c_2\sqrt{pm} + c_3\sqrt{\log n}\right) < \frac{c_4}{\sqrt{\log m}}. \tag{35}$$

Also, by direct calculation,

$$\boldsymbol{\Phi}_{v\times} = \frac{1}{\sqrt{m}}\boldsymbol{R}_v^*\boldsymbol{Q}_{C^*(v)\times}^*,$$

and it is a direct consequence that for all $v \in [n]$,

$$\left\|\boldsymbol{\Phi}_{v\times}\right\|_F = \frac{d}{m}. \tag{36}$$

Moreover, for v, u belonging to the same ground truth cluster, we have

$$\boldsymbol{\Phi}_{v\times}\boldsymbol{\Phi}_{u\times}^\top = \frac{1}{m}\boldsymbol{R}_v^*\boldsymbol{R}_u^{*\top}.$$

Lemma 8 then implies

$$\left\|\Pi_{\mathcal{O}(d)}\left(\widehat{\boldsymbol{U}}_{v\times}\widehat{\boldsymbol{U}}_{u\times}^\top\right) - \boldsymbol{R}_v^*\boldsymbol{R}_u^{*\top}\right\|_F \le 2m\left\|\widehat{\boldsymbol{U}}_{v\times}\widehat{\boldsymbol{U}}_{u\times}^\top - \boldsymbol{\Phi}_{v\times}\boldsymbol{\Phi}_{u\times}^\top\right\|_F. \tag{37}$$

Now we consider $u = \tau_i$ and $v \in \mathcal{I}_i^0 \cap \mathcal{I}_{\pi(i)}^*$. Suppose the following conditions hold for all $i \in [n]$, whose validity with high probability will be proved at the end of this section:

$$\begin{cases} \tau_i \in \mathcal{I}_{\pi(i)}^*; & (38)\\ \text{there exists a constant } c_5 > 0 \text{ such that } \left\|\widehat{\boldsymbol{U}}_{\tau_i\times} - \boldsymbol{\Phi}_{\tau_i\times}\right\|_F \le \sqrt{\frac{1}{(8Kd+c_5)\rho^2 m}}. & (39)\end{cases}$$

Then (37) yields

$$\left\|\boldsymbol{R}_v^0 - \boldsymbol{R}_v^*\boldsymbol{R}_{\tau_i}^{*\top}\right\|_F \le 2m\left\|\widehat{\boldsymbol{U}}_{v\times}\widehat{\boldsymbol{U}}_{\tau_i\times}^\top - \boldsymbol{\Phi}_{v\times}\boldsymbol{\Phi}_{\tau_i\times}^\top\right\|_F.$$

Therefore, if we denote $\tau_{C^*(v)} = \delta(v)$,

$$\begin{aligned}
&\sum_{i=1}^K \sum_{v\in\mathcal{I}_i^0\cap\mathcal{I}_{\pi(i)}^*} \left\|\boldsymbol{R}_v^0 - \boldsymbol{R}_v^*\boldsymbol{R}_{\tau_i}^{*\top}\right\|_F^2 \le 4m^2\sum_{v=1}^n \left\|\widehat{\boldsymbol{U}}_{v\times}\widehat{\boldsymbol{U}}_{\delta(v)\times}^\top - \boldsymbol{\Phi}_{v\times}\boldsymbol{\Phi}_{\delta(v)\times}^\top\right\|_F^2\\
\le& 8m^2\left(\sum_{v=1}^n \left\|\left(\widehat{\boldsymbol{U}}_{v\times} - \boldsymbol{\Phi}_{v\times}\right)\widehat{\boldsymbol{U}}_{\delta(v)\times}^\top\right\|_F^2 + \sum_{v=1}^n \left\|\left(\widehat{\boldsymbol{U}}_{\delta(v)\times} - \boldsymbol{\Phi}_{\delta(v)\times}\right)\boldsymbol{\Phi}_{v\times}^\top\right\|_F^2\right)\\
\le& 8m^2\max_{v\in[n]}\left\|\widehat{\boldsymbol{U}}_{\delta(v)\times}^\top\right\|_F^2 \times \sum_{v=1}^n\left\|\widehat{\boldsymbol{U}}_{v\times} - \boldsymbol{\Phi}_{v\times}\right\|_F^2 + 8m^2\frac{d}{m}\sum_{v=1}^n\left\|\widehat{\boldsymbol{U}}_{\delta(v)\times} - \boldsymbol{\Phi}_{\delta(v)\times}\right\|_F^2,
\end{aligned}$$

where the second inequality follows from triangle inequality, and the third from (36) and the inequality $\|\boldsymbol{X}\boldsymbol{Y}\|_F \le \|\boldsymbol{X}\|_F\|\boldsymbol{Y}\|_F$. Apply triangle inequality to (39), we have $\left\|\widehat{\boldsymbol{U}}_{\delta(v)\times}\right\|_F^2 \le \frac{c_6}{m}$ for some constant $c_6 > 0$. (39) again implies

$$\sum_{i=1}^K \sum_{v\in\mathcal{I}_i^0\cap\mathcal{I}_{\pi(i)}^*} \left\|\boldsymbol{R}_v^0 - \boldsymbol{R}_v^*\boldsymbol{R}_{\tau_i}^{*\top}\right\|_F^2 \le \frac{8c_4^2c_6m}{\log m} + 8m^2Kd\frac{1}{(8Kd+c_5)\rho^2m} < \frac{c_7m}{\rho^2},$$

where $0 < c_7 < 1$ is a constant. This yields

$$\begin{aligned}
\epsilon(\boldsymbol{V}^0)^2 &\le \frac{Cnd}{\log n} + \sum_{i=1}^K \sum_{v\in\mathcal{I}_i^0\cap\mathcal{I}_{\pi(i)}^*} \left\|\boldsymbol{R}_v^0 - \boldsymbol{R}_v^*\boldsymbol{R}_{\tau_i}^{*\top}\right\|_F^2\\
&\le \frac{Cnd}{\log n} + \frac{c_7m}{\rho^2} < \frac{m}{\rho^2},
\end{aligned}$$

which completes the proof.

We now show that (38) and (39) simultaneously hold with probability at least $1 - (\log n)^{-\Omega(1)}$. Firstly, according to (24), there exists a permutation π of the set $[K]$ such that

$$\left|\mathcal{I}_i^0 \cap \mathcal{I}_{\pi(i)}^*\right| \ge m - \frac{CKm}{2\log(Km)}$$

for arbitrary $i \in [K]$. Therefore, for any fixed i, τ_i picked in algorithm 3 satisfy (38) with probability at least

$$1 - \frac{CK}{2\log(Km)}. \tag{40}$$

Secondly, for any size-m set $T \subset [n]$, the size of the subset

$$\left\{ t \in T : \left\| \widehat{\boldsymbol{U}}_{t\times} - \boldsymbol{\Phi}_{t\times} \right\|_F^2 \leq \frac{1}{(8Kd + c_5)\rho^2 m} \right\}$$

is at least $m - \frac{m}{\sqrt{\log m}}$. Otherwise (35) is contradicted since

$$\left\| \widehat{\boldsymbol{U}} - \boldsymbol{\Phi} \right\|_F^2 \geq \sum_{t \in T} \left\| \widehat{\boldsymbol{U}}_{t\times} - \boldsymbol{\Phi}_{t\times} \right\|_F^2 > \frac{m}{\sqrt{\log m}} \frac{1}{(8Kd + c_5)\rho^2 m} > \frac{c_4^2}{\log m}$$

for a sufficiently large m. Therefore, for any fixed i, (39) holds with probability at least

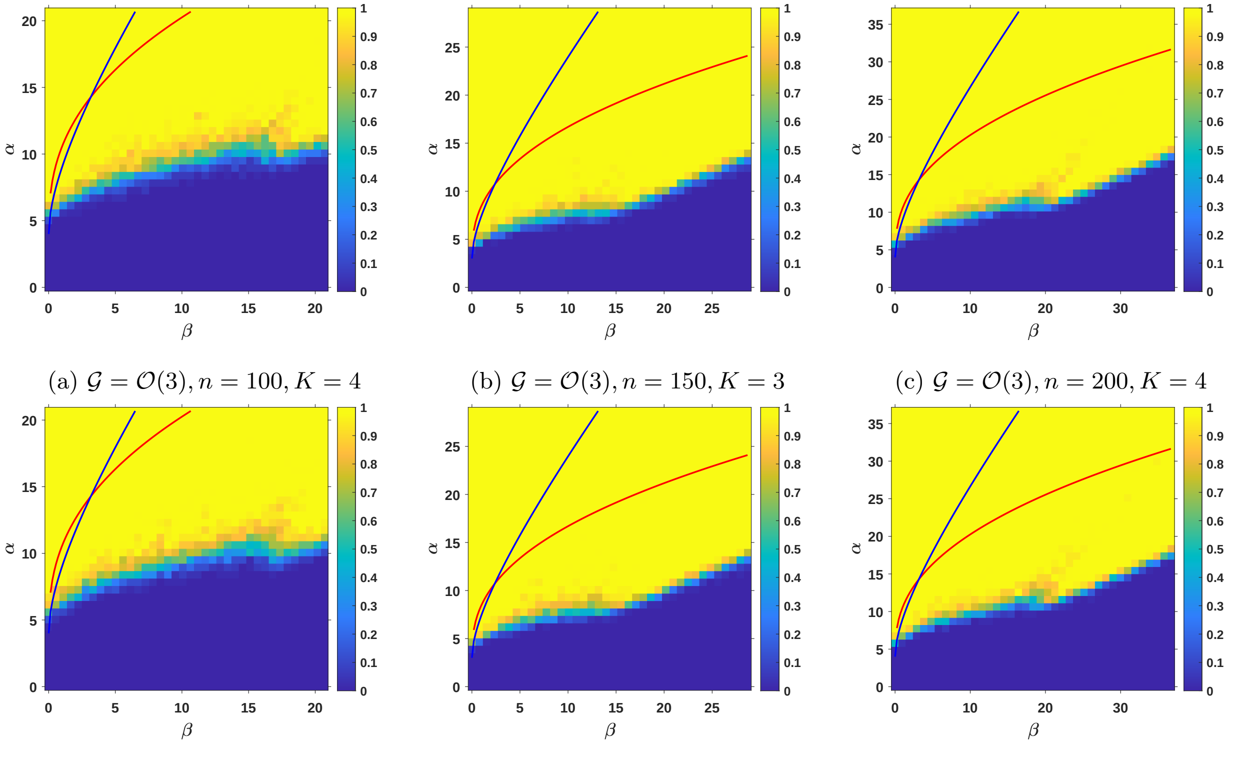
$$1 - \frac{1}{\sqrt{\log m}}. \tag{41}$$

By (40), (41) and union bound, (38) and (39) simultaneously hold for all i with probability at least

$$1 - K\left(\frac{CK}{2\log(Km)} + \frac{1}{\sqrt{\log m}} \right) = 1 - (\log n)^{-\Omega(1)}.$$

□

B ADDITIONAL EXPERIMENTS



(a) $\mathcal{G} = \mathcal{O}(3), n = 100, K = 4$

(b) $\mathcal{G} = \mathcal{O}(3), n = 150, K = 3$

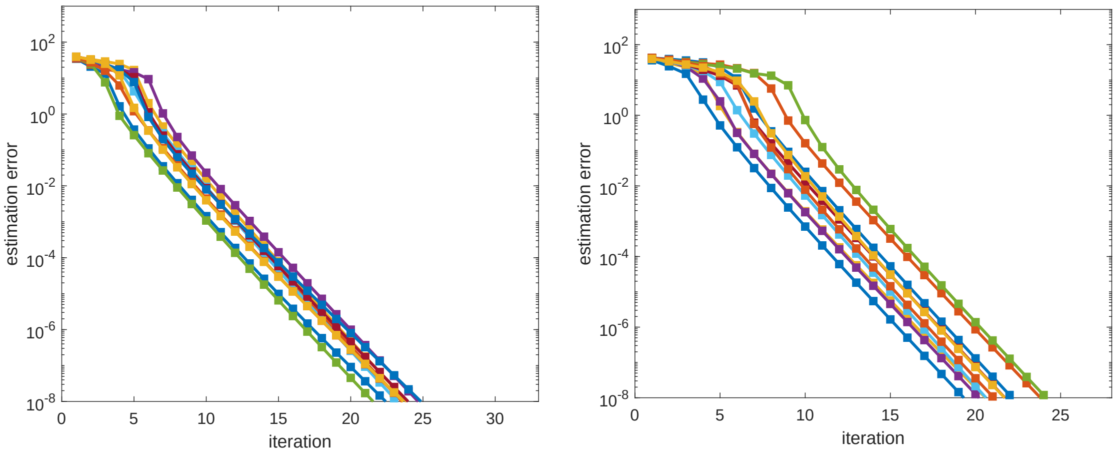
(c) $\mathcal{G} = \mathcal{O}(3), n = 200, K = 4$

(d) $\mathcal{G} = \mathcal{SO}(3), n = 100, K = 4$

(e) $\mathcal{G} = \mathcal{SO}(3), n = 150, K = 3$

(f) $\mathcal{G} = \mathcal{SO}(3), n = 200, K = 4$

Figure 4: Phase transition results on GPM with three different pairs of parameters $(n, K) = (100, 4)$, $(150, 3)$, and $(200, 4)$ in both orthogonal and rotational scenarios. The theoretical threshold for pure community detection $\sqrt{\alpha} - \sqrt{\beta} = \sqrt{K}$ is plotted in blue; the improved lower bound claimed in Theorem 2 is plotted in red.



(a) $K = 8, \alpha = 25, \beta = 15$

(b) $K = 10, \alpha = 35, \beta = 25$

Figure 5: Convergence results of GPM at $\mathcal{G} = \mathcal{O}(d)$ and $n = 400$, with parameters $(K, \alpha, \beta) = (8, 25, 15)$ and $(10, 35, 25)$.