The Annals of Statistics
2022, Vol. 50, No. 2, 1066–1094
https://doi.org/10.1214/21-AOS2140

ITERATIVE ALGORITHM FOR DISCRETE STRUCTURE RECOVERY

BY CHAO GAO[1,a] AND ANDERSON Y. ZHANG[2,b]

[1]*Department of Statistics, University of Chicago,* [a]*chaogao@uchicago.edu*
[2]*Department of Statistics, University of Pennsylvania,* [b]*ayz@wharton.upenn.edu*

We propose a general modeling and algorithmic framework for discrete structure recovery that can be applied to a wide range of problems. Under this framework, we are able to study the recovery of clustering labels, ranks of players, signs of regression coefficients, cyclic shifts and even group elements from a unified perspective. A simple iterative algorithm is proposed for discrete structure recovery, which generalizes methods including Lloyd's algorithm and the power method. A linear convergence result for the proposed algorithm is established in this paper under appropriate abstract conditions on stochastic errors and initialization. We illustrate our general theory by applying it on several representative problems: (1) clustering in Gaussian mixture model, (2) approximate ranking, (3) sign recovery in compressed sensing, (4) multireference alignment and (5) group synchronization, and show that minimax rate is achieved in each case.

1. Introduction. Discrete structure is commonly seen in modern statistics and machine learning, and various problems can be formulated into tasks of recovering the underlying discrete structure. A leading example is clustering analysis [51], where the discrete structure of the data is parametrized by a vector of clustering labels. Theoretical and algorithmic understandings of clustering analysis have received much attention in the recent literature especially due to the interest in community detection of network data [48, 57, 70, 91]. Other important examples of discrete structure recovery include ranking [20, 65], variable selection [22, 52], crowdsourcing [29, 41], estimation of unknown permutation [25, 74], graph matching [26, 32], and recovery of hidden Hamiltonian cycle [10, 21].

Despite the the progress of understanding discrete structures in various specific problems, a general theoretical investigation has been lacking in the literature. This is partly due to the fact that theory of discrete structure recovery can be quite different from traditional statistical estimation of continuous parameters. In fact, it has been argued that the nature of discrete structure recovery is closely related to hypothesis testing theory [42]. In addition, the existing literature on the statistical guarantees of discrete structure recovery mostly focuses on characterizing the condition of exact recovery [1, 10, 61, 66, 69, 84, 93]. Let $z^* = (z_1^*, z_2^*, \ldots, z_p^*)$ represent a discrete structure of interest, where each z_j^* parametrizes a discrete status of either the jth sample or the jth variable of the data set. The exact recovery is achieved by some estimator $\widehat{z}$ if $\widehat{z}_j = z_j^*$ for all $j \in [p]$. However, exact recovery of discrete structure usually requires a strong signal to noise ratio condition. A more interesting, more realistic, but harder problem is when only partial recovery [22, 40, 41, 44, 72, 90, 91] of z^* is possible. Under this regime, a statistical guarantee can be established on the proportion of errors, and the result will naturally lead to the condition of exact recovery as a special case.

Discrete structure recovery is also challenging from a computational point of view. In spite of being optimal in many cases, maximum likelihood estimation of z^* is often combinatorial, and thus computationally infeasible. Though convex relaxations such as linear programming

Received September 2020; revised September 2021.
MSC2020 subject classifications. 62F07.
Key words and phrases. k-means clustering, approximate ranking, high-dimensional statistics, multireference alignment, group synchronization.

or semidefinite programming can be derived for many specific problems [10, 13, 47, 49, 50, 58, 67], they may not be scalable to very large data sets and the analysis of partial recovery of convex relaxation is usually quite involved [34, 35, 47]. Moreover, in many examples such as clustering and variable selection, the data generating process is parametrized both by a discrete structure and a continuous model parameter. The presence of the nuisance continuous parameter further complicates the design of efficient algorithms.

The goal of this paper is to develop a general modeling and algorithmic framework for partial recovery of discrete structures. We first propose a general structured linear model parametrized by a discrete structure z^* and a global continuous parameter B^*, which unifies various problems of discrete structure recovery into the same framework. A simple iterative algorithm is then proposed for recovering z^*, which can be informally written in the following form:

$$z^{(t)} = \operatorname*{argmin}_{z} \sum_{j=1}^{p} \|T_j - \nu_j(\widehat{B}(z^{(t-1)}), z_j)\|^2 \quad \text{for all } t \geq 1. \tag{1}$$

Here, T_j is some local statistic whose distribution depends both on the jth label z_j^* and the global continuous parameter B^* of the model. Because of the separability of the objective function across $j \in [p]$, each $z_j^{(t)}$ takes the value of z_j such that $\nu_j(\widehat{B}(z^{(t-1)}), z_j)$ is the closest to T_j and, therefore, computation of (1) is straightforward. The general iterative procedure (1) recovers some interesting algorithms, among which perhaps the most important one is Lloyd's algorithm [60] for k-means clustering. In the clustering context, T_j is the jth data point, and $\nu_j(\widehat{B}(z^{(t-1)}), z_j)$ is the z_jth estimated clustering center computed based on the clustering labels $z^{(t-1)}$ from the previous step. In addition, (1) also leads to algorithms in approximate ranking, sign recovery and many other problems that will be studied in details in this paper.

The main result of our paper characterizes conditions under which (1) converges with respect to some loss function $\ell(\cdot, \cdot)$ to be defined later. An informal statement of the result is given:

$$\ell(z^{(t)}, z^*) \leq 2\xi_{\text{ideal}}(\delta) + \frac{1}{2}\ell(z^{(t-1)}, z^*) \quad \text{for all } t \geq 1, \tag{2}$$

with high probability. That is, the value of $\ell(z^{(t)}, z^*)$ converges at a linear rate to $4\xi_{\text{ideal}}(\delta)$. Here, we use $\xi_{\text{ideal}}(\delta)$ to characterize the error of an ideal procedure,

$$\widehat{z}^{\text{ideal}} = \operatorname*{argmin}_{z} \sum_{j=1}^{p} \|T_j - \nu_j(\widehat{B}(z^*), z_j)\|^2, \tag{3}$$

and the definition of $\xi_{\text{ideal}}(\delta)$ with a general $\delta > 0$ will be given in Section 3. The convergence result (2) is established with some $\delta > 0$ arbitrarily close to 0. We note that the ideal procedure (3) is not realizable because of its dependence on the true z^*, but (2) shows that the iterative algorithm (1) achieves almost the same statistical performance of (3). The general abstract result is then applied to several concrete examples: clustering for Gaussian mixture model, approximate ranking, sign recovery in compressed sensing, multireference alignment and group synchronization, which represent different types of discrete structure recovery problems. Moreover, in each of the examples, we can relate $\xi_{\text{ideal}}(\delta)$ to the minimax rate of the problem and, therefore, claim that the simple algorithm (1) is both computationally efficient and minimax optimal.

Another popular method that is suitable for discrete structure recovery is the EM algorithm [30]. The global convergence of EM algorithm has been established under the setting of unimodal likelihood [85] and the setting of two-component Gaussian mixtures [28, 86–88].

Local convergence results for general settings are obtained by [11]. However, the most important difference between [11] and our work, besides the obvious difference of algorithms, is that our convergence guarantee (2) is established for the estimation error of the discrete structure z^*, while the convergence result in [11] for the EM algorithm is established for the estimation error of the continuous model parameter B^*. Results like (2) may be possibly established for the EM algorithm in the context of clustering using the techniques suggested by the paper [92],[1] but whether (2) can be proved for the EM algorithm in general settings is unknown.

The most related work to us in the literature is the analysis of Lloyd's algorithm in Gaussian mixture models by [62]. Since Lloyd's algorithm is a special case of (1), our convergence result (2) recovers the result in [62] as a special case with even a slightly weaker condition on the number of clusters. We also mention the recent paper [71] that studies a variant of Lloyd's algorithm and improves the signal to noise ratio condition in [62] for the two-component Gaussian mixtures.

Organization. Our general modeling and algorithmic framework will be introduced in Section 2. In Section 3, we formulate abstract conditions under which we can establish the convergence of the algorithm. Applications to specific examples will be discussed afterwards, including clustering in Gaussian mixture model (Section 4), sign recovery in compressed sensing (Section 5), multireference alignment (Section 6) and group synchronization (Section 7). Section 8 discusses the potential limitations of our framework and possible open problems. The application to approximate ranking and all the technical proofs will be given in the Supplementary Material [46].

Notation. For $d \in \mathbb{N}$, we write $[d] = \{1, \dots, d\}$. Given $a, b \in \mathbb{R}$, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For two positive sequences a_n and b_n, we write $a_n \lesssim b_n$ to mean that there exists a constant $C > 0$ independent of n such that $a_n \le C b_n$ for all n; moreover, $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a set S, we use $\mathbb{I}\{S\}$ and $|S|$ to denote its indicator function and cardinality, respectively. For a vector $v = (v_1, \dots, v_d)^T \in \mathbb{R}^d$, we define $\|v\|^2 = \sum_{\ell=1}^d v_\ell^2$. The trace inner product between two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$ is defined as $\langle A, B \rangle = \sum_{\ell=1}^{d_1} \sum_{\ell'=1}^{d_2} A_{\ell\ell'} B_{\ell\ell'}$, while the Frobenius and operator norms of A are given by $\|A\|_{\mathrm{F}} = \sqrt{\langle A, A \rangle}$ and $\|A\|_{\mathrm{op}} = s_{\max}(A)$, respectively, where $s_{\max}(\cdot)$ denotes the largest singular value. The notation $\mathbb{P}$ and $\mathbb{E}$ are generic probability and expectation operators whose distribution is determined from the context.

2. A general framework of models and algorithms. We start with the introduction of structured linear model. Consider a pair of random vectors $Y \in \mathbb{R}^N$ and $X \in \mathbb{R}^D$. We impose the relation that

$$\mathbb{E}(Y|X) = \mathscr{X}_{z^*}(B^*). \tag{4}$$

On the right-hand side of (4), $z^* = (z_1^*, \dots, z_p^*)$ is a vector of discrete labels, and each z_j^* is allowed to take its value from a label set of size k. For simplicity, we assume the label set to be $[k]$ without loss of generality. The vector B^* is the model parameter that lives in a subspace indexed by z^*. We use the notation $\mathcal{B}_{z^*}$ for this subspace. Finally, $\mathscr{X}_{z^*}$ is a linear operator jointly determined by X and z^*. It maps from $\mathcal{B}_{z^*}$ to $\mathbb{R}^N$.

[1]The paper [92] established the convergence of mean-field coordinate ascent and Gibbs sampling in the sense of (2) for community detection in stochastic block models. Due to the connection and similarity between the EM algorithm and variational Bayes, we believe the techniques used in (2) can also be applied to the analysis of EM algorithms for clustering problems.

The general structured linear model (4) can be viewed as a slight variation of the one introduced by [45]. It is particularly suitable for the research of label recovery and includes some important examples that will be studied in this paper.

To estimate the labels $z_1^*, \dots, z_p^*$, one strategy is to first compute a local statistic $T_j = T_j(X, Y) \in \mathbb{R}^d$ and then infer z_j^* from T_j for each $j \in [p]$. We require that

$$\mathbb{E}(T_j|X) = \mu_j\big(B^*, z_j^*\big). \tag{5}$$

Then suppose the model parameter B^* was known, a natural procedure to estimate z_j^* would find an $a \in [k]$ such that $\|T_j - \mu_j(B^*, a)\|^2$ is the smallest. However, for some applications, the form of $\mu_j(B^*, z_j^*)$ may not be available, and thus we need to associate each $\mu_j(B^*, z_j^*)$ with a surrogate $\nu_j(B^*, z_j^*)$. An oracle procedure that uses the knowledge of B^* is given by

$$\widehat{z}_j^{\text{oracle}} = \operatorname*{argmin}_{a\in[k]} \big\|T_j - \nu_j\big(B^*, a\big)\big\|^2. \tag{6}$$

On the other hand, since B^* is unknown in practice, we need to replace the B^* in (6) by an estimator. A natural procedure is the least-squares estimator $\widehat{B}(z^*)$, where for a given z, $\widehat{B}(z)$ is defined by

$$\widehat{B}(z) = \operatorname*{argmin}_{B\in\mathcal{B}_z} \big\|Y - \mathscr{X}_z(B)\big\|^2. \tag{7}$$

This time we need to know z in (7) to compute $\widehat{B}(z)$. Therefore, we shall combine (6) and (7) and obtain the following iterative algorithm.

Algorithm 1: Iterative discrete structure recovery

Input : The data Y, X and the number of iterations $t_{\max}$.

Output: The estimator $\widehat{z} = z^{(t_{\max})}$.

1 Compute the initializer $z^{(0)}$.

2 For t in $1 : t_{\max}$, compute

$$B^{(t)} = \operatorname*{argmin}_{B\in\mathcal{B}_{z^{(t-1)}}} \big\|Y - \mathscr{X}_{z^{(t-1)}}(B)\big\|^2 \quad \text{and} \tag{8}$$

$$z_j^{(t)} = \operatorname*{argmin}_{a\in[k]} \big\|T_j(X, Y) - \nu_j\big(B^{(t)}, a\big)\big\|^2 \quad \forall j \in [p]. \tag{9}$$

Let us now discuss a few important examples. Though we regard X and Y to be vectors in our general framework, in some specific examples, it is often more convenient to arrange the data into matrices instead of vectors. Of course, the two representations are equivalent and the relation can be precisely described with the operations of vectorization and Kronecker product.

2.1. *Clustering in Gaussian mixture model.* Consider $Y \in \mathbb{R}^{d\times p}$ with $Y_1, \dots, Y_p$ standing for its columns. We assume that $Y_j \sim \mathcal{N}(\theta_{z_j^*}^*, I_d)$ independently for $j \in [p]$. Here, $z_1^*, \dots, z_p^* \in [k]$ are p clustering labels and $\theta_1^*, \dots, \theta_k^* \in \mathbb{R}^d$ are k clustering centers. In our general framework, we have $N = dp$, B^* is the concatenation of the k clustering centers, and $\mathcal{B}_{z^*} = \mathbb{R}^{d\times k}$. The linear operator $\mathscr{X}_{z^*}$ maps the matrix $\{\theta_a^*\}_{a\in[k]} \in \mathbb{R}^{d\times k}$ to the matrix $\{\theta_{z_j^*}^*\}_{j\in[p]} \in \mathbb{R}^{d\times p}$. For the algorithm to recover the clustering labels, the obvious local statistic is $T_j = Y_j$ for $j \in [p]$. Moreover, we set $\nu_j(B^*, a) = \mu_j(B^*, a) = \theta_a^*$. Then Algorithm 1

is specialized into the following iterative procedures:

$$\theta_a^{(t)} = \frac{\sum_{j=1}^p \mathbb{I}\{z_j^{(t-1)} = a\} Y_j}{\sum_{j=1}^p \mathbb{I}\{z_j^{(t-1)} = a\}}, \quad a \in [k],$$
$$z_j^{(t)} = \operatorname*{argmin}_{a \in [k]} \|Y_j - \theta_a^{(t)}\|^2, \quad j \in [p].$$

This is recognized as Lloyd's algorithm [60], the most popular way to solve k-means clustering.

2.2. *Approximate ranking.* In the task of ranking, we consider the observation of pairwise interaction data Y_{ij} for $(i, j) \in [p]^2$ and $i \neq j$. The rank or the position of the jth player is specified by an integer $z_j^* \in [p]$. What is known as the pairwise comparison model assumes that $Y_{ij} \sim \mathcal{N}(\beta^*(z_i^* - z_j^*), 1)$ for some signal strength parameter $\beta^* \in \mathbb{R}$. Our goal is to estimate the discrete position z_j^* for each player $j \in [p]$. This is known as the approximate ranking problem [40], which is different from exact ranking where z^* corresponds to a permutation. It is easy to see that this approximate ranking model is a special case of our general structured linear model. To be specific, we have $N = p(p-1)$, B^* is identified with β^*, and $\mathcal{B}_{z^*} = \mathbb{R}$. The linear operator $\mathscr{X}_{z^*}$ maps β^* to $\{\beta^*(z_i^* - z_j^*)\}_{1 \le i \neq j \le p}$. To recover z_j^*, it is natural to define

$$T_j = \frac{1}{\sqrt{2(p-1)}} \sum_{i \in [p] \setminus \{j\}} (Y_{ji} - Y_{ij}). \tag{10}$$

Thus, we have

$$\mu_j(B^*, a) = \frac{2p}{\sqrt{2(p-1)}} \beta^* \left(a - \frac{1}{p} \sum_{i=1}^p z_i^* \right). \tag{11}$$

Because of the dependence of $\mu_j(B^*, a)$ on the unknown $\frac{1}{p}\sum_{i=1}^p z_i^*$, we also introduce $\nu_j(B^*, a)$ that replaces $\frac{1}{p}\sum_{i=1}^p z_i^*$ with a fixed value $\frac{p+1}{2}$,

$$\nu_j(B^*, a) = \frac{2p}{\sqrt{2(p-1)}} \beta^* \left(a - \frac{p+1}{2} \right).$$

The choice of $\frac{p+1}{2}$ is due to the parameter space of z^* that will be made specific in the Supplementary Material [46]. This leads to the following iterative algorithm:

$$\begin{aligned} \beta^{(t)} &= \frac{\sum_{1 \le i \neq j \le p} (z_i^{(t-1)} - z_j^{(t-1)}) Y_{ij}}{\sum_{1 \le i \neq j \le p} (z_i^{(t-1)} - z_j^{(t-1)})^2}, \\ z_j^{(t)} &= \operatorname*{argmin}_{a \in [p]} \left| \sum_{i \in [p] \setminus \{j\}} (Y_{ji} - Y_{ij}) - 2p\beta^{(t)} \left(a - \frac{p+1}{2} \right) \right|^2, \quad j \in [p]. \end{aligned} \tag{12}$$

Since (12) is recognized as feature matching [25], this is the iterative feature matching algorithm suggested by [40] for approximate ranking. The statistical property of the above algorithm will be analyzed in the Supplementary Material [46] due to page limit.

2.3. *Sign recovery in compressed sensing.* In a standard regression problem, we assume $Y|X \sim \mathcal{N}(X\beta^*, I_n)$. Consider a random design setting, where $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ for $(i, j) \in [n] \times [p]$. We study the sign recovery problem, which is equivalent to estimating

$z_j^* \in \{-1, 0, 1\}$, where the three possible values of z_j^* standing for β_j^* being negative, zero and positive. We also define the sparsity level $s = \sum_{j=1}^p |z_j^*|$. In order that sign recovery is information-theoretically possible, we assume that either $\beta_j^* = 0$ or $|\beta_j^*| \geq \lambda$. The same setting has been considered by [72]. The sparse linear regression model is clearly a special case of our general framework with the choices $N = n$, $B^* = \beta^*$ and $\mathcal{B}_{z^*} = \{\beta \in \mathbb{R}^p : \beta_j = \beta_j |z_j^*|\}$. The linear operator $\mathscr{X}_{z^*}$ maps β^* to $X\beta^*$. Following [72], we use the local statistic

$$T_j = \|X_j\|^{-1} X_j^T Y \tag{13}$$

to recover z_j^*. Here, $X_j \in \mathbb{R}^n$ stands for the jth column of X. Computing its conditional expectation, we obtain

$$\mu_j(B^*, a) = a\|X_j\| \max\{|\beta_j^*|, \lambda\} + \|X_j\|^{-1} \sum_{l \in [p]\setminus\{j\}} \beta_l^* X_j^T X_l, \tag{14}$$

for $a \in \{-1, 0, 1\}$, because of the assumption that β_j^* is either 0 or larger than λ in magnitude. Replacing $\max\{|\beta_j^*|, \lambda\}$ in the above formula by some threshold level $2t(X_j)$, we get

$$v_j(B^*, a) = 2a\|X_j\| t(X_j) + \|X_j\|^{-1} \sum_{l \in [p]\setminus\{j\}} \beta_l^* X_j^T X_l, \tag{15}$$

for $a \in \{-1, 0, 1\}$. The threshold level is specified by

$$t(X_j) = \frac{\lambda}{2} + \frac{\log \frac{p-s}{s}}{\lambda \|X_j\|^2}, \tag{16}$$

which can be derived from a minimax analysis [22, 72]. Specializing Algorithm 1 to the current context gives

$$\beta^{(t)} = \operatorname*{argmin}_{\{\beta \in \mathbb{R}^p : \beta_j = \beta_j |z_j^{(t-1)}|\}} \|y - X\beta\|^2, \tag{17}$$

$$z_j^{(t)} = \begin{cases} 1, & \dfrac{X_j^T Y - \sum_{l \in [p]\setminus\{j\}} \beta_l^{(t)} X_j^T X_l}{\|X_j\|^2} > t(X_j), \\ 0, & -t(X_j) \leq \dfrac{X_j^T Y - \sum_{l \in [p]\setminus\{j\}} \beta_l^{(t)} X_j^T X_l}{\|X_j\|^2} \leq t(X_j), \\ -1, & \dfrac{X_j^T Y - \sum_{l \in [p]\setminus\{j\}} \beta_l^{(t)} X_j^T X_l}{\|X_j\|^2} < -t(X_j). \end{cases} \tag{18}$$

We note that (18) is a slight modification of the variable selection procedure in [72]. The main difference is that [72] uses an estimator of β^* computed with an independent data set, while we compute a least-squares procedure (17) restricted on the support of $z^{(t-1)}$ obtained from the previous step using the same data set.

2.4. *Multireference alignment.* Consider independent data points $Y_j \sim \mathcal{N}(Z_j^*\theta^*, I_d)$ for $j \in [p]$. A common parameter $\theta^* \in \mathbb{R}^d$ is shared by the p observations. The matrix Z_j^* is a cyclic shift such that $(Z_j\theta^*)_i = \theta^*_{i+t_j \,(\mathrm{mod}\, d)}$ for some integer t_j. In other words, for each $j \in [p]$, a noisy shifted version is observed. To put the problem into our general framework, we have $N = dp$, $B^* = \theta^*$, $z_j^* = Z_j^*$, and $\mathcal{B}_{z_j^*} = \mathbb{R}^d$. The linear operator $\mathscr{X}_{z^*}$ maps θ^* to $(Z_1^*\theta^*, \ldots, Z_p^*\theta^*)$. We are interested in the recovery of the cyclic shifts $Z_1^*, \ldots, Z_p^*$. For this purpose, consider the local statistic $T_j = Y_j$ for all $j \in [p]$. This results in $v_j(B^*, U) =$

$\mu_j(B^*, U) = U\theta^*$ for any $U \in \mathcal{C}_d$ where $\mathcal{C}_d$ is the class of cyclic shifts. Then Algorithm 1 is specialized into the following iterative procedures:

$$\theta^{(t)} = \frac{1}{p}\sum_{j=1}^{p}(Z_j^{(t-1)})^T Y_j,$$

$$Z_j^{(t)} = \operatorname*{argmin}_{U\in\mathcal{C}_d} \|Y_j - U\theta^{(t)}\|^2, \quad j \in [p].$$

It is clear that $|\mathcal{C}_d| = d$, and thus the update for $Z_j^{(t)}$ has a linear complexity.

2.5. *Group synchronization.* Consider a group $(\mathcal{G}, \circ)$ and group elements $g_1, \dots, g_p \in \mathcal{G}$. The group synchronization problem is the recovery of $g_1, \dots, g_p$ from noisy versions of $g_i \circ g_j^{-1}$. It turns out a number of important instances of group synchronization can be regarded as special cases of our general framework, and thus can be provably solved by the iterative algorithm. Let us illustrate by the simplest example of $\mathbb{Z}_2$ synchronization. In this model, one observes $Y_{ij} \sim \mathcal{N}(\lambda^* z_i^* z_j^*, 1)$ for all $1 \le i < j \le p$ with $z_1^*, \dots, z_p^* \in \{-1, 1\}$. The parameter $\lambda^* \in \mathbb{R}$ plays the role of signal-to-noise ratio.

Though it is most natural to identify λ^* with β^* in the general framework, this treatment would result in an iterative algorithm with a computationally infeasible update of $z^{(t)}$ because of the quadratic dependence. A smart and much better way is to regard the vector $\lambda^* z^* \in \mathbb{R}^p$ as β^* by taking advantage of the flexibility of $\mathcal{B}_{z^*}$. To be specific, we can organize the observations into a matrix $Y = z^*(\beta^*)^T + W \in \mathbb{R}^{p\times p}$ with $\beta^* \in \mathcal{B}_{z^*} = \{\beta = \lambda z^* : \lambda \in \mathbb{R}\}$. In this way, we have $N = \frac{p(p-1)}{2}$, $B^* = \beta^*$, and $\mathcal{B}_{z^*} = \{\beta = \lambda z^* : \lambda \in \mathbb{R}\}$. Thus, given β^*, the mean of Y is linear with respect to z^*. To derive an iterative algorithm, we let $T_j = Y_j$ be the jth column of Y. We then have $\nu_j(B^*, a) = \mu_j(B^*, a) = a\beta^*$ for $a \in \{-1, 1\}$. The iterative algorithm is

$$\begin{aligned} \beta^{(t)} &= \operatorname*{argmin}_{\beta=\lambda z^{(t-1)}:\lambda\in\mathbb{R}} \|Y - z^{(t-1)}\beta^T\|_{\mathrm{F}}^2, \\ z_j^{(t)} &= \operatorname*{argmin}_{a\in\{-1,1\}} \|Y_j - a\beta^{(t)}\|^2. \end{aligned} \tag{19}$$

It is easy to see that (19) has a closed form $\beta^{(t)} = \frac{(z^{(t-1)})^T Y z^{(t-1)}}{p^2} z^{(t-1)}$. This leads to the following equivalent form of the algorithm:

$$z_j^{(t)} = \begin{cases} \operatorname{sign}(Y_j^T z^{(t-1)}), & (z^{(t-1)})^T Y z^{(t-1)} \ge 0, \\ -\operatorname{sign}(Y_j^T z^{(t-1)}), & (z^{(t-1)})^T Y z^{(t-1)} < 0, \end{cases} \tag{20}$$

which is a variation of the power method.

In addition to $\mathbb{Z}_2$ synchronization, the above idea can also be applied to other group synchronization problems. Examples of $\mathbb{Z}/k\mathbb{Z}$ synchronization and permutation synchronization will be analyzed in the Supplementary Material [46].

3. Convergence analysis. In this section, we formulate abstract conditions under which we can derive the statistical and computational guarantees of Algorithm 1.

3.1. *A general loss function.* Our goal is to establish a bound for every $t \ge 1$ with respect to the loss $\ell(z^{(t)}, z^*)$. The loss function is defined by

$$\ell(z, z^*) = \sum_{j=1}^{p} \|\mu_j(B^*, z_j) - \mu_j(B^*, z_j^*)\|^2. \tag{21}$$

It has a close relation to the Hamming loss $h(z, z^*) = \sum_{j=1}^{p} \mathbb{I}\{z_j \neq z_j^*\}$. Define

$$\Delta_{\min}^2 = \min_{j\in[p]} \min_{1\le a\neq b\le k} \|\mu_j(B^*, a) - \mu_j(B^*, b)\|^2,$$

and then we immediately have

$$\ell(z, z^*) \ge \Delta_{\min}^2 h(z, z^*). \tag{22}$$

3.2. *Error decomposition.* By (5), we can decompose each local statistic as

$$T_j = \mu_j(B^*, z_j^*) + \epsilon_j. \tag{23}$$

We usually have $\epsilon_j \sim \mathcal{N}(0, I_d)$, but this is not required, and we shall also note that the ϵ_j's may not even be independent across $j \in [p]$. By (9), if we start the algorithm from any z, then z_j^* will be incorrectly estimated after one iteration if $z_j^* \neq \operatorname{argmin}_{a\in[k]} \|T_j - \nu_j(\widehat{B}(z), a)\|^2$. Consequently, assume $z_j^* = a$, and it is important to analyze the event

$$\|T_j - \nu_j(\widehat{B}(z), b)\|^2 \le \|T_j - \nu_j(\widehat{B}(z), a)\|^2, \tag{24}$$

for any $b \in [k]\backslash\{a\}$. Recall the definition of $\widehat{B}(z)$ in (7). We plug (23) into (24), and then after some rearrangement, we can see that the event (24) is equivalent to

$$\begin{aligned} &\langle \epsilon_j, \nu_j(\widehat{B}(z^*), a) - \nu_j(\widehat{B}(z^*), b)\rangle \\ &\quad\le -\frac{1}{2}\Delta_j(a, b)^2 + F_j(a, b; z) + G_j(a, b; z) + H_j(a, b). \end{aligned} \tag{25}$$

On the right-hand side of (25), $\Delta_j(a, b)^2$ is the main term that characterizes the difference between the two labels a and b. It is defined as

$$\Delta_j(a, b)^2 = \|\mu_j(B^*, a) - \nu_j(B^*, b)\|^2 - \|\mu_j(B^*, a) - \nu_j(B^*, a)\|^2.$$

Note that with the notation $\Delta_j(a, b)^2$, we have implicitly assume that $\Delta_j(a, b)^2 \ge 0$ throughout the paper. This assumption is easily satisfied in all of the examples considered in the paper. The other three terms in (25) are the error terms that we need to control. Their definitions are given by

$$\begin{aligned} F_j(a, b; z) &= \langle \epsilon_j, (\nu_j(\widehat{B}(z^*), a) - \nu_j(\widehat{B}(z), a)) - (\nu_j(\widehat{B}(z^*), b) - \nu_j(\widehat{B}(z), b))\rangle, \\ G_j(a, b; z) &= \frac{1}{2}(\|\mu_j(B^*, a) - \nu_j(\widehat{B}(z), a)\|^2 - \|\mu_j(B^*, a) - \nu_j(\widehat{B}(z^*), a)\|^2) \\ &\quad - \frac{1}{2}(\|\mu_j(B^*, a) - \nu_j(\widehat{B}(z), b)\|^2 - \|\mu_j(B^*, a) - \nu_j(\widehat{B}(z^*), b)\|^2), \\ H_j(a, b) &= \frac{1}{2}(\|\mu_j(B^*, a) - \nu_j(\widehat{B}(z^*), a)\|^2 - \|\mu_j(B^*, a) - \nu_j(B^*, a)\|^2) \\ &\quad - \frac{1}{2}(\|\mu_j(B^*, a) - \nu_j(\widehat{B}(z^*), b)\|^2 - \|\mu_j(B^*, a) - \nu_j(B^*, b)\|^2). \end{aligned}$$

With these quantities defined as above, we can check that (25) is indeed equivalent to (24). To help readers understand the meaning of these error terms, we work out the formulas in the context of $\mathbb{Z}_2$ synchronization. By specializing the definitions of the error terms in $\mathbb{Z}_2$ synchronization, we have for any $a \neq b$,

$$F_j(a, b; z) = (a - b)\langle \epsilon_j, \widehat{\beta}(z^*) - \widehat{\beta}(z)\rangle, \tag{26}$$

$$G_j(a, b; z) = 2\langle \beta^*, \widehat{\beta}(z^*) - \widehat{\beta}(z)\rangle, \tag{27}$$

$$H_j(a, b) = 2\langle \beta^*, \beta^* - \widehat{\beta}(z^*)\rangle. \tag{28}$$

The reason to have such decomposition (25) is as follows:

- By igoring the three error terms, the event $\langle \epsilon_j, \nu_j(\widehat{B}(z^*), a) - \nu_j(\widehat{B}(z^*), b)\rangle \le -\frac{1}{2}\Delta_j(a, b)^2$ contributes to the ideal recovery error rate. That is, even if we were given the true z^*, applying one iteration in Algorithm 1, that is, (9) would still result in some error.
- The error terms $F_j(a, b; z)$ and $G_j(a, b; z)$ can be controlled by the difference between $\widehat{B}(z)$ and $\widehat{B}(z^*)$, which further depends on $\ell(z, z^*)$. We will treat $F_j(a, b; z)$ and $G_j(a, b; z)$ differently because the former involves the additional randomness of ϵ_j.
- The error term $H_j(a, b)$ can be controlled by the difference between $\widehat{B}(z^*)$ and B^*. In fact, unlike $F_j(a, b; z)$ or $G_j(a, b; z)$, $H_j(a, b)$ does not depend on z, and thus its value remains unchanged throughout the iterations.

3.3. *Conditions for algorithmic convergence.* Now we need to discuss how to analyze the error terms $F_j(a, b; z)$, $G_j(a, b; z)$ and $H_j(a, b)$. There are three types of conditions that we will impose.

CONDITION A (ℓ_2-type error control). Assume that

$$\max_{\{z:\ell(z,z^*)\le\tau\}} \sum_{j=1}^{p} \max_{b\in[k]\setminus\{z_j^*\}} \frac{F_j(z_j^*, b; z)^2 \|\mu_j(B^*, b) - \mu_j(B^*, z_j^*)\|^2}{\Delta_j(z_j^*, b)^4 \ell(z, z^*)} \le \frac{1}{256}\delta^2$$

holds with probability at least $1 - \eta_1$, for some $\tau, \delta, \eta_1 > 0$.

CONDITION B (Restricted ℓ_2-type error control). Assume that

$$\max_{\{z:\ell(z,z^*)\le\tau\}} \max_{T\subset[p]} \frac{\tau}{4\Delta_{\min}^2|T| + \tau} \sum_{j\in T} \max_{b\in[k]\setminus\{z_j^*\}} \frac{G_j(z_j^*, b; z)^2 \|\mu_j(B^*, b) - \mu_j(B^*, z_j^*)\|^2}{\Delta_j(z_j^*, b)^4 \ell(z, z^*)}$$

$$\le \frac{1}{256}\delta^2$$

holds with probability at least $1 - \eta_2$, for some $\tau, \delta, \eta_2 > 0$.

CONDITION C (ℓ_∞-type error control). Assume that

$$\max_{j\in[p]} \max_{b\in[k]\setminus\{z_j^*\}} \frac{|H_j(z_j^*, b)|}{\Delta_j(z_j^*, b)^2} \le \frac{1}{4}\delta$$

holds with probability at least $1 - \eta_3$, for some $\tau, \delta, \eta_3 > 0$.

Conditions A, B and C are for the error terms $F_j(a, b; z)$, $G_j(a, b; z)$ and $H_j(a, b)$, respectively. Because of the difference of the three terms that we have mentioned earlier, they are controlled in different ways. Both Conditions A and B impose ℓ_2-type controls and relate $F_j(a, b; z)$ and $G_j(a, b; z)$ to the loss function $\ell(z, z^*)$. On the other hand, $H_j(a, b)$ is controlled by an ℓ_∞-type bound in Condition C.

Next, we define a quantity referred to as the ideal error,

$$\begin{aligned}\xi_{\text{ideal}}(\delta) = {} & \sum_{j=1}^{p} \sum_{b\in[k]\setminus\{z_j^*\}} \|\mu_j(B^*, b) - \mu_j(B^*, z_j^*)\|^2 \\ & \times \mathbb{I}\Big\{\langle \epsilon_j, \nu_j(\widehat{B}(z^*), z_j^*) - \nu_j(\widehat{B}(z^*), b)\rangle \le -\frac{1-\delta}{2}\Delta_j(z_j^*, b)^2\Big\}.\end{aligned} \tag{29}$$

We note that $\xi_{\text{ideal}}(\delta)$ is a quantity that does not change with t. In fact, with some $\delta > 0$, $\xi_{\text{ideal}}(\delta)$ can be shown to be an error bound for the ideal procedure $\widehat{z}_j^{\text{ideal}}$ defined in (3). We

therefore choose $\xi_{\text{ideal}}(\delta)$ with a small $\delta > 0$ as the target error that $z^{(t)}$ converges to. In specific examples studied later in Sections 4–5, we will show $\xi_{\text{ideal}}(\delta)$ can be bounded by the minimax rate of each problem.

CONDITION D (Ideal error). Assume that

$$\xi_{\text{ideal}}(\delta) \le \frac{1}{4}\tau, \tag{30}$$

with probability at least $1 - \eta_4$, for some $\tau, \delta, \eta_4 > 0$.

Finally, we need a condition on $z^{(0)}$, the initialization of Algorithm 1.

CONDITION E (Initialization). Assume that

$$\ell\big(z^{(0)}, z^*\big) \le \tau,$$

with probability at least $1 - \eta_5$, for some $\tau, \eta_5 > 0$.

3.4. *Convergence guarantee.* With all the conditions specified, we establish the convergence guarantee for Algorithm 1.

THEOREM 3.1. *Assume Conditions* A, B, C, D *and* E *hold for some* $\tau, \delta, \eta_1, \eta_2, \eta_3, \eta_4, \eta_5 > 0$. *We then have*

$$\ell\big(z^{(t)}, z^*\big) \le 2\xi_{\text{ideal}}(\delta) + \frac{1}{2}\ell\big(z^{(t-1)}, z^*\big) \quad \textit{for all } t \ge 1,$$

with probability at least $1 - \eta$, *where* $\eta = \sum_{i=1}^{5} \eta_i$.

The theorem shows that the error of $z^{(t)}$ converges to $4\xi_{\text{ideal}}(\delta)$ at a linear rate. Among all the conditions, Conditions A, B and C are the most important ones. The largest τ that makes Conditions A, B and C hold simultaneously will be the required error bound for the initialization in Condition E. With (22), Theorem 3.1 also implies that the iterative algorithm achieves an error of $4\xi_{\text{ideal}}(\delta)/\Delta_{\min}^2$ in terms of Hamming distance.

In Sections 4–7, we will apply Theorem 3.1 to the examples mentioned in Section 2, covering different categories of discrete structures: clustering label, rank, variable sign, cyclic shift and group element. The clustering labels are discrete objects without order or any topological structure. This is in contrast to the ranks that are ordered objects in the space of natural numbers. The variable signs are similar to the clustering labels except two differences. The first difference is the prior knowledge that most variables are zero in the context of sparse linear regression. The second difference is that a nonzero sign only implies a range of a variable instead of its specific value. Group elements have their own unique properties that depend on the specific settings. Despite all the differences between these discrete structures, we are able to analyze them in a unified framework with the same algorithm.

4. Clustering in Gaussian mixture model. We assume the data matrix $Y \in \mathbb{R}^{d\times p}$ is generated from a Gaussian mixture model. This means we have $Y_j = \theta_{z_j^*} + \epsilon_j \sim \mathcal{N}(\theta_{z_j^*}, I_d)$ independently for $j \in [p]$, where $z^* \in [k]^p$ is the vector of clustering labels that we aim to recover. Specializing Algorithm 1 to the clustering problem, we obtain the well-known Lloyd's algorithm, which can be summarized as

$$z_j^{(t)} = \operatorname*{argmin}_{a\in[k]} \big\|Y_j - \widehat{\theta}_a\big(z^{(t-1)}\big)\big\|^2, \quad j \in [p],$$

where for each $z \in [k]^p$, we use the notation

$$\widehat{\theta}_a(z) = \frac{\sum_{j=1}^p \mathbb{I}\{z_j = a\} Y_j}{\sum_{j=1}^p \mathbb{I}\{z_j = a\}}, \quad a \in [k].$$

Even though general k-means clustering is known to be NP-hard [7, 27, 64], local convergence of the Lloyd's iteration can be established under certain data-generating mechanism [9, 54]. In particular, the recent work [62] shows that under the Gaussian mixture model, the misclustering error of $z^{(t)}$ in the Lloyd's iteration linearly converges to the minimax optimal rate. In this section, we show that our theoretical framework developed in Section 3 leads to a result that is comparable to the one in [62].

4.1. *Conditions.* To analyze the algorithmic convergence, we note that $\mu_j(B^*, a) = \nu_j(B^*, a) = \theta_a^*$, $\Delta_j(a,b)^2 = \|\theta_a^* - \theta_b^*\|^2$, $\ell(z, z^*) = \sum_{j=1}^p \|\theta_{z_j}^* - \theta_{z_j^*}^*\|^2$, and $\Delta_{\min} = \min_{1 \le a \ne b \le k} \|\theta_a^* - \theta_b^*\|$ in the current setting. The error terms that we need to control are

$$F_j(a,b;z) = \langle \epsilon_j, \widehat{\theta}_a(z^*) - \widehat{\theta}_a(z) - \widehat{\theta}_b(z^*) + \widehat{\theta}_b(z) \rangle,$$

$$G_j(a,b;z) = \frac{1}{2}\big(\|\theta_a^* - \widehat{\theta}_a(z)\|^2 - \|\theta_a^* - \widehat{\theta}_a(z^*)\|^2 - \|\theta_a^* - \widehat{\theta}_b(z)\|^2 + \|\theta_a^* - \widehat{\theta}_b(z^*)\|^2\big),$$

$$H_j(a,b) = \frac{1}{2}\big(\|\theta_a^* - \widehat{\theta}_a(z^*)\|^2 - \|\theta_a^* - \widehat{\theta}_b(z^*)\|^2 + \|\theta_a^* - \theta_b^*\|^2\big).$$

The following lemma controls the error terms $F_j(a,b;z)$, $G_j(a,b;z)$ and $H_j(a,b)$.

LEMMA 4.1. *Assume that* $\min_{a \in [k]} \sum_{j=1}^p \mathbb{I}\{z_j^* = a\} \ge \frac{\alpha p}{k}$ *and* $\tau \le \frac{\Delta_{\min}^2 \alpha p}{2k}$ *for some constant* $\alpha > 0$. *Then, for any* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* α *and* C' *such that*

$$\max_{\{z:\ell(z,z^*)\le\tau\}} \sum_{j=1}^p \max_{b \in [k]\setminus\{z_j^*\}} \frac{F_j(z_j^*, b; z)^2 \|\mu_j(B^*, b) - \mu_j(B^*, z_j^*)\|^2}{\Delta_j(z_j^*, b)^4 \ell(z, z^*)}$$
$$\le C \frac{k^2(kd/p + 1)}{\Delta_{\min}^2}\Big(1 + \frac{k(d/p+1)}{\Delta_{\min}^2}\Big), \tag{31}$$

$$\max_{\{z:\ell(z,z^*)\le\tau\}} \max_{T \subset [p]} \frac{\tau}{4\Delta_{\min}^2 |T| + \tau} \sum_{j \in T} \max_{b \in [k]\setminus\{z_j^*\}} \frac{G_j(z_j^*, b; z)^2 \|\mu_j(B^*, b) - \mu_j(B^*, z_j^*)\|^2}{\Delta_j(z_j^*, b)^4 \ell(z, z^*)}$$
$$\le C\Big(\frac{k\tau}{p\Delta_{\min}^2} + \frac{k(d+p)}{p\Delta_{\min}^2} + \frac{k^2(d+p)^2}{p^2\Delta_{\min}^4}\Big), \tag{32}$$

and

$$\max_{j \in [p]} \max_{b \in [k]\setminus\{z_j^*\}} \frac{|H_j(z_j^*, b)|}{\Delta_j(z_j^*, b)^2} \le C\Big(\frac{k(d + \log p)}{p\Delta_{\min}^2} + \sqrt{\frac{k(d+\log p)}{p\Delta_{\min}^2}}\Big), \tag{33}$$

with probability at least $1 - p^{-C'}$.

From the bounds (31)–(33), we can see that a sufficient condition that Conditions A, B and C hold is $\frac{\tau}{p\Delta_{\min}^2/k} \to 0$ and

$$\frac{\Delta_{\min}^2}{k^2(kd/p+1)} \to \infty. \tag{34}$$

In fact, under this sufficient condition, we can set $\delta = \delta_p$ to be some sequence δ_p converging to 0 in Conditions A, B and C.

Next, we need to control $\xi_{\text{ideal}}(\delta)$ in Condition D. This is given by the following lemma.

LEMMA 4.2. *Assume* $\frac{\Delta_{\min}^2}{\log k + kd/p} \to \infty$, $p/k \to \infty$, *and* $\min_{a\in[k]} \sum_{j=1}^{p} \mathbb{I}\{z_j^* = a\} \geq \frac{\alpha p}{k}$ *for some constant* $\alpha > 0$. *Then, for any sequence* $\delta_p = o(1)$, *we have*

$$\xi_{\text{ideal}}(\delta_p) \leq p \exp\biggl(-\bigl(1 + o(1)\bigr)\frac{\Delta_{\min}^2}{8}\biggr),$$

with probability at least $1 - \exp(-\Delta_{\min})$.

We note that the signal condition $\frac{\Delta_{\min}^2}{\log k + kd/p} \to \infty$ required by Lemma 4.2 is implied by the stronger condition (34). Therefore, we need to require (34) for the Conditions A, B, C and D to hold simultaneously.

4.2. *Convergence.* With the help of Lemma 4.1 and Lemma 4.2, we can specialize Theorem 3.1 into the following result.

THEOREM 4.1. *Assume* (34) *holds*, $p/k \to \infty$, *and* $\min_{a\in[k]} \sum_{j=1}^{p} \mathbb{I}\{z_j^* = a\} \geq \frac{\alpha p}{k}$ *for some constant* $\alpha > 0$. *Suppose* $z^{(0)}$ *satisfies*

$$\ell\bigl(z^{(0)}, z^*\bigr) = o\biggl(\frac{p\Delta_{\min}^2}{k}\biggr), \tag{35}$$

with probability at least $1 - \eta$. *Then we have*

$$\ell\bigl(z^{(t)}, z^*\bigr) \leq p \exp\biggl(-\bigl(1 + o(1)\bigr)\frac{\Delta_{\min}^2}{8}\biggr) + \frac{1}{2}\ell\bigl(z^{(t-1)}, z^*\bigr) \quad \textit{for all } t \geq 1,$$

with probability at least $1 - \eta - \exp(-\Delta_{\min}) - p^{-1}$.

REMARK 4.1. Our result is comparable to the main result of [62]. The main difference is that the convergence analysis in [62] is for the misclustering error, defined by

$$\mathsf{Misclust}(z, z^*) = \frac{1}{p}\sum_{j=1}^{p} \mathbb{I}\bigl\{z_j \neq z_j^*\bigr\}, \tag{36}$$

while Theorem 4.1 is established for an ℓ_2 type loss function, which is more natural in our general framework. The main condition of Theorem 4.1 is the signal requirement (34). Interestingly, this is exactly the same condition used in [62]. On the other hand, we only require $k = o(p)$ for the number of clusters allowed, whereas [62] assumes a slightly stronger condition $k = o(p/(\log p)^{1/3})$.

In the context of clustering, the loss function (36) may be more natural than $\ell(z, z^*)$. Given the relation that

$$\mathsf{Misclust}(z, z^*) \leq \frac{\ell(z, z^*)}{p\Delta_{\min}^2},$$

we immediately obtain the following corollary on the misclustering error.

COROLLARY 4.1. *Assume* (34) *holds*, $p/k \to \infty$, *and* $\min_{a\in[k]} \sum_{j=1}^{p} \mathbb{I}\{z_j^* = a\} \geq \frac{\alpha p}{k}$ *for some constant* $\alpha > 0$. *Suppose* $z^{(0)}$ *satisfies* (35) *with probability at least* $1-\eta$. *Then we have*

$$\mathsf{Misclust}\big(z^{(t)}, z^*\big) \leq \exp\bigg(-\big(1+o(1)\big)\frac{\Delta_{\min}^2}{8}\bigg) + 2^{-t} \quad \textit{for all } t \geq 1, \tag{37}$$

with probability at least $1-\eta-\exp(-\Delta_{\min})-p^{-1}$.

According to a lower bound result in [62], the quantity $\exp(-(1+o(1))\frac{\Delta_{\min}^2}{8})$ is the minimax rate of recovering z^* with respect to the loss function $\mathsf{Misclust}(z, z^*)$ under the Gaussian mixture model. Since $\mathsf{Misclust}(z, z^*)$ takes value in the set $\{j/p : j \in [p] \cup \{0\}\}$, the term 2^{-t} in (37) is negligible as long as $2^{-t} = o(p^{-1})$. We therefore can claim

$$\mathsf{Misclust}\big(z^{(t)}, z^*\big) \leq \exp\bigg(-\big(1+o(1)\big)\frac{\Delta_{\min}^2}{8}\bigg) \quad \text{for all } t \geq 3\log p.$$

In other words, the minimax rate is achieved after at most $\lceil 3\log p \rceil$ iterations.

4.3. *Initialization*. To close this section, we discuss how to initialize Lloyd's algorithm. In the literature, this is usually done by spectral methods [9, 54, 62]. We consider the following variation that is particularly suitable for Gaussian mixture models. Our initialization procedure has two steps:

1. Perform a singular value decomposition on Y, and obtain $Y = \sum_{l=1}^{p\wedge n} \widehat{d}_l \widehat{u}_l \widehat{v}_l^T$ with $\widehat{d}_1 \geq \cdots \geq \widehat{d}_{p\wedge n} \geq 0$, $\{\widehat{u}_l\}_{l\in[p\wedge n]} \in \mathbb{R}^d$ and $\{\widehat{v}_l\}_{l\in[p\wedge n]} \in \mathbb{R}^p$. With $\widehat{U} = (\widehat{u}_1, \ldots, \widehat{u}_k) \in \mathbb{R}^{d\times k}$, we define

$$\widehat{\mu} = \widehat{U}^T Y \in \mathbb{R}^{k\times p}. \tag{38}$$

2. Find some $\beta_1^{(0)}, \ldots, \beta_k^{(0)} \in \mathbb{R}^k$ and $z^{(0)} \in [k]^p$ that satisfy

$$\sum_{j=1}^{p} \big\|\widehat{\mu}_j - \beta_{z_j^{(0)}}^{(0)}\big\|^2 \leq M \min_{\substack{\beta_1,\ldots,\beta_k\in\mathbb{R}^k \\ z\in[k]^p}} \sum_{j=1}^{p} \|\widehat{\mu}_j - \beta_{z_j}\|^2, \tag{39}$$

where $\widehat{\mu}_j$ is the jth column of $\widehat{\mu}$.

The first step (38) serves as a dimensionality reduction procedure, which reduces the dimension of data from d to k. Then the columns of $\widehat{\mu}$ are collected to compute the M-approximation of the k-means objective in (39). We note that approximation of the k-means objective can be computed efficiently in polynomial time [8, 53, 55]. For example, the k-means++ algorithm [8] can efficiently solve (39) with $M = O(\log k)$. However, we shall treat M flexible here, and its value will be reflected in the error bound of $z^{(0)}$. The second step (39) can also be replaced by a greedy clustering algorithm used in [43]. The theoretical guarantee of $z^{(0)}$ is given in the following proposition.

PROPOSITION 4.1. *Assume* $\min_{a\in[k]} \sum_{j=1}^{p} \mathbb{I}\{z_j^* = a\} \geq \frac{\alpha p}{k}$ *for some constant* $\alpha > 0$ *and* $\Delta_{\min}^2/((M+1)k^2(1+d/p)) \to \infty$. *For any* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* α *and* C' *such that*

$$\min_{\pi\in\Pi_k} \ell\big(\pi \circ z^{(0)}, z^*\big) \leq C(M+1)k(p+d), \tag{40}$$

with probability at least $1-e^{-C'(p+d)}$, *where* Π_k *denotes the set of permutations on* $[k]$.

We remark that a signal to noise ratio condition that is sufficient for both the conclusions of Proposition 4.1 and Theorem 4.1 is given by

$$\frac{\Delta_{\min}^2}{(M+1)k^2(kd/p+1)} \to \infty, \tag{41}$$

which is almost identical to (34). Note that the clustering structure is only identifiable up to a label permutation, and this explains the necessity of the minimum over Π_k in (40). In other words, (40) implies that there exists some $\pi \in \Pi_k$, such that $\ell(z^{(0)}, \pi^{-1} \circ z^*) \le C(M+1)k^2(p+d)$. Then, under the condition (41), (35) is satisfied with z^* replaced by $\pi^{-1} \circ z^*$. Therefore, Theorem 4.1 implies that $\ell(z^{(t)}, \pi^{-1} \circ z^*)$ converges to the minimax error with a linear rate.

5. Sign recovery in compressed sensing. We consider a regression model $Y = X\beta^* + w \in \mathbb{R}^n$, where $X \in \mathbb{R}^{n\times p}$ is a random design matrix with i.i.d. entries $X_{ij} \sim \mathcal{N}(0,1)$, and w is an independent noise vector with i.i.d. entries $w_i \sim \mathcal{N}(0,1)$. Our goal is to recover the signs of the regression coefficients β_j^*'s. Formally speaking, we assume

$$z^* \in \mathcal{Z}_s = \Big\{ z \in \{-1,0,1\}^p : \sum_{j=1}^p |z_j| = s \Big\},$$

and $\beta^* \in \mathcal{B}_{z^*,\lambda}$, where for some $z \in \{-1,0,1\}^p$ and some $\lambda > 0$, the space $\mathcal{B}_{z,\lambda}$ is defined by

$$\mathcal{B}_{z,\lambda} = \Big\{ \beta \in \mathbb{R}^p : \beta_j = z_j|\beta_j|, \min_{\{j\in[p]: z_j \neq 0\}} |\beta_j| \ge \lambda \Big\}.$$

The problem is to estimate the sign vector z^*. A closely related problem is support recovery, which is equivalent to estimating the vector $\{|z_j^*|\}_{j\in[p]}$. This problem has received much attention in the literature of compressed sensing, where one usually has control over the distribution of the design matrix. Necessary and sufficient conditions on (n, p, s, λ) for exact support recovery have been derived in [5, 36, 77, 78, 82, 83] and references therein. Recently, the minimax rate of partial support recovery with respect to the Hamming loss has been derived in [72]. Their results can be easily modified to the estimation of the sign vector z^* as well. We will state the lower bound result in [72] as our benchmark. To do that, we need to introduce the normalized Hamming loss

$$\mathsf{H}_{(s)}(z, z^*) = \frac{1}{s} h(z, z^*) = \frac{1}{s} \sum_{j=1}^p \mathbb{I}\{z_j \neq z_j^*\}.$$

We also define the signal-to-noise ratio of the problem by

$$\mathsf{SNR} = \frac{\lambda\sqrt{n}}{2} - \frac{\log \frac{p-s}{s}}{\lambda\sqrt{n}}. \tag{42}$$

THEOREM 5.1 (Ndaoud and Tsybakov [72]). *Assume* $\limsup s/p < \frac{1}{2}$ *and* $s \log p \le n$. *If* $\mathsf{SNR} \to \infty$, *we have*

$$\inf_{\widehat{z}} \sup_{z^*\in\mathcal{Z}_s} \sup_{\beta^*\in\mathcal{B}_{z^*,\lambda}} \mathbb{E}\mathsf{H}_{(s)}(\widehat{z}, z^*) \ge \exp\Big(-\frac{(1+o(1))\mathsf{SNR}^2}{2}\Big) - 4e^{-s/8}.$$

Otherwise, if $\mathsf{SNR} = O(1)$, *we then have*

$$\inf_{\widehat{z}} \sup_{z^*\in\mathcal{Z}_s} \sup_{\beta^*\in\mathcal{B}_{z^*,\lambda}} \mathbb{E}\mathsf{H}_{(s)}(\widehat{z}, z^*) \ge c,$$

for some constant $c > 0$.

We remark that the lower bound result in [72] is stated in a more general nonasymptotic form. Here, we choose to work out its asymptotic formula (by Lemma G.2) so that we can better compare the lower bound with the upper bound rate achieved by our algorithm. In [72], the minimax rate is achieved by a thresholding procedure that requires sample splitting. Though theoretically sound, the requirement of splitting the data into two halves may not be appealing in practice. This is where our general Algorithm 1 comes. We will show that Algorithm 1 can achieve the minimax rate without sample splitting.

Our analysis is focused in the regime where $\mathsf{SNR} \to \infty$, which is necessary for consistency under the loss $\mathsf{H}_{(s)}(\widehat{z}, z^*)$ according to Theorem 5.1. Specializing Algorithm 1 to the current setting, we obtain the following iterative procedure:

$$z_j^{(t)} = \begin{cases} 1, & \dfrac{X_j^T Y - \sum_{l\in[p]\setminus\{j\}} \widehat{\beta}_l(z^{(t-1)}) X_j^T X_l}{\|X_j\|^2} > t(X_j), \\ 0, & -t(X_j) \le \dfrac{X_j^T Y - \sum_{l\in[p]\setminus\{j\}} \widehat{\beta}_l(z^{(t-1)}) X_j^T X_l}{\|X_j\|^2} \le t(X_j), \\ -1, & \dfrac{X_j^T Y - \sum_{l\in[p]\setminus\{j\}} \widehat{\beta}_l(z^{(t-1)}) X_j^T X_l}{\|X_j\|^2} < -t(X_j) \end{cases} \quad j \in [p], \tag{43}$$

where $t(X_j)$ is defined by (16). Here, for some $z \in \{-1, 0, 1\}^p$, we use the notation

$$\widehat{\beta}(z) = \operatorname*{argmin}_{\{\beta\in\mathbb{R}^p : \beta_j = \beta_j |z_j|\}} \|y - X\beta\|^2.$$

In other words, $\widehat{\beta}(z)$ is the least-squares solution on the support of z. The formula (43) resembles the thresholding procedure proposed in [72]. In [72], $\widehat{\beta}_l(z^{(t-1)})$ is replaced by some estimator $\widehat{\beta}_l$ computed from an independent data set. In comparison, we use $\widehat{\beta}_l(z^{(t-1)})$, and thus avoid sample splitting. The iteration (43) is also different from existing algorithms in the literature for support/sign recovery in compressed sensing. For example, the popular iterative hard thresholding algorithm [17] updates the regression coefficients with a gradient step instead of a full least-squares step. The hard thresholding pursuit algorithm [37] has a full least-squares steps, but updates the support by choosing the s variables with the largest absolute values.

5.1. *Conditions.* For any $j \in [p]$, T_j is the local statistic defined in (13) and it can be decomposed as $T_j = \mu_j(B^*, z_j^*) + \epsilon_j$, with $\epsilon_j = \|X_j\|^{-1} X_j^T w \sim \mathcal{N}(0, 1)$. To analyze the algorithmic convergence, we need to specialize the abstract objects $\|\mu_j(B^*, z_j^*) - \mu_j(B^*, b)\|^2$, $\Delta_j(z_j^*, b)^2$, and $\ell(z, z^*)$ into the current setting. With the formulas (14) and (15), we have

$$\|\mu_j(B^*, z_j^*) - \mu_j(B^*, b)\|^2 = \begin{cases} \lambda^2 \|X_j\|^2 m & z_j^* = 0 \text{ and } b \ne 0 m \\ |\beta_j^*|^2 \|X_j\|^2, & z_j^* \ne 0 \text{ and } b = 0, \\ 4|\beta_j^*|^2 \|X_j\|^2, & z_j^* b = -1, \end{cases} \tag{44}$$

which leads to the formula of the loss function

$$\ell(z, z^*) = \sum_{j=1}^p \big(\lambda^2 \|X_j\|^2 \mathbb{I}\{z_j^* = 0, z_j \ne 0\} + |\beta_j^*|^2 \|X_j\|^2 \mathbb{I}\{z_j^* \ne 0, z_j = 0\} + 4|\beta_j^*|^2 \|X_j\|^2 \mathbb{I}\{z_j z_j^* = -1\}\big).$$

By (22), we have the relation

$$\mathsf{H}_{(s)}(z, z^*) \le \frac{\ell(z, z^*)}{s\Delta_{\min}^2}, \tag{45}$$

where $\Delta_{\min}^2 = \lambda^2 \min_{j\in[p]} \|X_j\|^2$ in the current setting. Lastly, the formula of $\Delta_j(z_j^*, b)^2$ is given by

$$\Delta_j(z_j^*, b)^2 = \begin{cases} 4t(X_j)^2\|X_j\|^2, & z_j^* = 0 \text{ and } b \neq 0, \\ 4t(X_j)(|\beta_j^*| - t(X_j))\|X_j\|^2, & z_j^* \neq 0 \text{ and } b = 0, \\ 8t(X_j)|\beta_j^*|\|X_j\|^2, & z_j^* b = -1. \end{cases} \tag{46}$$

One may question whether we always have $\Delta_j(z_j^*, b)^2 > 0$ for all $b \neq z_j^*$ and $j \in [p]$. We note that this property is guaranteed by Lemma G.3 with high probability.

Next, we analyze the error terms. In the current setting, they are

$$\begin{aligned} F_j(a, b; z) &= 0, \\ G_j(a, b; z) &= 2(a-b)t(X_j) \sum_{l\in[p]\setminus\{j\}} (\widehat{\beta}_l(z) - \widehat{\beta}_l(z^*))X_j^T X_l, \\ H_j(a, b) &= 2(a-b)t(X_j) \sum_{l\in[p]\setminus\{j\}} (\widehat{\beta}_l(z^*) - \beta^*)X_j^T X_l. \end{aligned}$$

LEMMA 5.1. *Assume $s \log p \le n$ and $\tau \le C_0 sn\lambda^2$ for some constant $C_0 > 0$. Then, for any $C' > 0$, there exists a constant $C > 0$ only depending on C_0 and C' such that*

$$\begin{aligned} &\max_{\{z:\ell(z,z^*)\le\tau\}} \max_{T\subset[p]} \frac{\tau}{4\Delta_{\min}^2|T| + \tau} \\ &\quad\times \sum_{j\in T} \max_{b\in\{-1,0,1\}\setminus\{z_j^*\}} \frac{G_j(z_j^*, b; z)^2\|\mu_j(B^*, b) - \mu_j(B^*, z_j^*)\|^2}{\Delta_j(z_j^*, b)^4 \ell(z, z^*)} \\ &\quad\le C\frac{s(\log p)^2}{n}\left(1 + \frac{1}{n\lambda^2}\right) \max_{j\in[p]}\left[\frac{|\beta_j^*|^2}{(|\beta_j^*| - t(X_j))^2} \vee \frac{\lambda^2}{t(X_j)^2}\right], \end{aligned} \tag{47}$$

and

$$\max_{j\in[p]} \max_{b\in\{-1,0,1\}\setminus\{z_j^*\}} \frac{|H_j(z_j^*, b)|}{\Delta_j(z_j^*, b)^2} \le C\sqrt{\frac{s(\log p)^2}{n}} \frac{1}{\min_{j\in[p]} \sqrt{n}||\beta_j^*| - t(X_j)|}, \tag{48}$$

with probability at least $1 - p^{-C'}$.

The two error bounds (47) and (48) are complicated. However, by Lemma G.3, if we additionally assume $\limsup s/p < \frac{1}{2}$, $\mathsf{SNR} \to \infty$, and $s(\log p)^4 = o(n)$, the right-hand sides of (47) and (48) can be shown to be of order $o((\log p)^{-1})$. Therefore, Conditions A, B and C hold with some $\delta = \delta_p = o((\log p)^{-1})$.

The following lemma controls $\xi_{\text{ideal}}(\delta)$ in Condition D.

LEMMA 5.2. *Assume $\limsup s/p < \frac{1}{2}$, $s \log p \le n$, and $\mathsf{SNR} \to \infty$. Then for any sequence $\delta_p = o((\log p)^{-1})$, we have*

$$\xi_{\text{ideal}}(\delta_p) \le sn\lambda^2 \exp\left(-\frac{(1 + o(1))\mathsf{SNR}^2}{2}\right),$$

with probability at least $1 - \exp(-\mathsf{SNR}) - p^{-1}$.

5.2. *Convergence.* With Lemma 5.1 and Lemma 5.2, we then can specialize Theorem 3.1 into the following result.

THEOREM 5.2. *Assume* $\limsup s/p < \frac{1}{2}$, $s(\log p)^4 = o(n)$, *and* $\mathsf{SNR} \to \infty$. *Suppose* $\ell(z^{(0)}, z^*) \le C_0 sn\lambda^2$ *with probability at least* $1-\eta$ *for some constant* $C_0 > 0$. *Then we have*

$$\ell\big(z^{(t)}, z^*\big) \le sn\lambda^2 \exp\bigg(-\frac{(1+o(1))\mathsf{SNR}^2}{2}\bigg) + \frac{1}{2}\ell\big(z^{(t-1)}, z^*\big) \quad \textit{for all } t \ge 1,$$

with probability at least $1-\eta-\exp(-\mathsf{SNR})-2p^{-1}$.

The relation (45) and a simple concentration result for $\min_{j\in[p]}\|X_j\|^2$ immediately implies a convergence result for the loss $\mathsf{H}_{(s)}(z, z^*)$.

COROLLARY 5.1. *Assume* $\limsup s/p < \frac{1}{2}$, $s(\log p)^4 = o(n)$, *and* $\mathsf{SNR} \to \infty$. *Suppose* $\ell(z^{(0)}, z^*) \le C_0 sn\lambda^2$ *with probability at least* $1-\eta$ *for some constant* $C_0 > 0$. *Then we have*

$$\mathsf{H}_{(s)}\big(z^{(t)}, z^*\big) \le \exp\bigg(-\frac{(1+o(1))\mathsf{SNR}^2}{2}\bigg) + 2^{-t} \quad \textit{for all } t \ge 1, \tag{49}$$

with probability at least $1-\eta-\exp(-\mathsf{SNR})-2p^{-1}$.

Since the loss function $\mathsf{H}_{(s)}(z, z^*)$ takes value in the set $\{j/s : j \in [p] \cap \{0\}\}$, the term 2^{-t} in (49) is negligible as long as $2^{-t} = o(s^{-1})$. We therefore can claim

$$\mathsf{H}_{(s)}\big(z^{(t)}, z^*\big) \le \exp\bigg(-\frac{(1+o(1))\mathsf{SNR}^2}{2}\bigg) \quad \text{for all } t \ge 3\log s,$$

when $s \to \infty$. If instead we have $s = O(1)$, then any $t \to \infty$ will do. This implies after at most $\lceil 3\log p\rceil$ iterations, Algorithm 1 achieves the minimax rate.

REMARK 5.1. The leading term of the nonasymptotic minimax lower bound in [72] with respect to the loss $\mathsf{H}_{(s)}(z, z^*)$ takes the form of $\psi(n, p, s, \lambda, 0)/s$, where

$$\psi(n, p, s, \lambda, \delta) = s\mathbb{P}\big(\epsilon > (1-\delta)\|\zeta\|\big(\lambda - t(\zeta)\big)\big) + (p-s)\mathbb{P}\big(\epsilon > (1-\delta)\|\zeta\|t(\zeta)\big) \tag{50}$$

with $\epsilon \sim \mathcal{N}(0,1)$ and $\zeta \sim \mathcal{N}(0, I_n)$ independent of each other. By scrutinizing the proof of Lemma 5.2, we can also write (49) as

$$\mathsf{H}_{(s)}\big(z^{(t)}, z^*\big) \lesssim \psi(n, p, s, \lambda, \delta_p)/s + 2^{-t} \quad \text{for all } t \ge 1,$$

with high probability with some $\delta_p = o((\log p)^{-1})$.

5.3. *Initialization.* Our final task in this section is to provide an initialization procedure that satisfies the bound $\ell(z^{(0)}, z^*) \le C_0 sn\lambda^2$ with high probability. We consider a simple procedure that thresholds the solution of the square-root SLOPE [15, 18, 31, 81]. It has the following two steps:

1. Compute

$$\widetilde{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\big(\|Y - X\beta\| + A\|\beta\|_{\mathrm{SLOPE}}\big), \tag{51}$$

where the penalty takes the form $\|\beta\|_{\mathrm{SLOPE}} = \sum_{j=1}^p \sqrt{\log(2p/j)}|\beta|_{(j)}$. Here $|\beta|_{(1)} \ge |\beta|_{(2)} \ge \cdots \ge |\beta|_{(p)}$ is a nonincreasing ordering of $|\beta_1|, |\beta_2|, \ldots, |\beta_p|$.

2. For any $j \in [p]$, compute $z_j^{(0)} = \operatorname{sign}(\widetilde{\beta}_j)\mathbb{I}\{|\widetilde{\beta}_j| \ge \lambda/2\}$.

The theoretical guarantee of $z^{(0)}$ is given by the following proposition.

PROPOSITION 5.1. *Assume* $\limsup s/p < \frac{1}{2}$, $s \log p \le n$, *and* $\mathsf{SNR} \to \infty$. *For some sufficiently large constant* $A > 0$ *in* (51) *and any constant* $C' > 0$, *there exist some* C_0 *and* C_1 *only depending on* A *and* C', *such that*

$$\ell\big(z^{(0)}, z^*\big) \le C_0 sn\lambda^2,$$

with probability at least $1 - e^{-C_1 s \log(ep/s)} - p^{-C'}$.

6. Multireference alignment. Multireference alignment [6, 13] is an important problem in mathematical chemistry and captures fundamental aspects of applications such as cryogenic electron microscopy (cryo-EM) [38, 79]. In this problem, we have observations $Y_j \sim \mathcal{N}(Z_j^*\theta^*, I_d)$ for $j \in [p]$ with $\theta^* \in \mathbb{R}^d$ and $Z_j^* \in \mathcal{C}_d$. Here, $\mathcal{C}_d \subset \{0,1\}^{d\times d}$ is the set of index cyclic shifts. For any $U \in \mathcal{C}_d$, there exists some integer t such that $(Uv)_i = v_{i+t \,(\mathrm{mod}\, d)}$ for any vector $v \in \mathbb{R}^d$. The literature for this problem has been focused on the recovery of the common signal parameter $\theta^* \in \mathbb{R}^d$ shared by the p observations [2, 13, 14, 16, 68, 75]. To the best of our knowledge, optimal estimation of the cyclic shifts $Z_1^*, \dots, Z_p^*$ still remains an open problem, and this is the focus of the current section. As is already discussed in Section 2, the cyclic shifts can be recovered by the general iterative algorithm. Before giving the statistical guarantee of the algorithm, we first present a minimax lower bound of the problem. We define

$$\Delta_{\min}^2 = \min_{U\in\mathcal{C}_d} \big\|(I_d - U)\theta^*\big\|^2.$$

This quantity plays the role of the minimal signal strength of the problem, and is very different from the signal strength required for estimating θ^* in [75]. Note that $\Delta_{\min}^2$ is a function of θ^*, and thus captures the difficulty of the problem for each instance of $\theta^* \in \mathbb{R}^d$. For example, if each coordinate of θ^* takes the same value, the corresponding $\Delta_{\min}^2 = 0$, and thus it is impossible to recover the shifts. The quantity $\Delta_{\min}^2$ is thus a characterization of the diversity of the sequence $\{\theta_i^*\}$.

THEOREM 6.1. *If* $\Delta_{\min}^2 \to \infty$, *we have*

$$\inf_{\widehat{Z}} \sup_{Z^*} \mathbb{E} \min_{U\in\mathcal{C}_d} \frac{1}{p}\sum_{j=1}^p \mathbb{I}\{\widehat{Z}_j U \ne Z_j^*\} \ge \exp\bigg(-\frac{(1+o(1))\Delta_{\min}^2}{8}\bigg).$$

Otherwise, if $\Delta_{\min}^2 = O(1)$, *we then have*

$$\inf_{\widehat{Z}} \sup_{Z^*} \mathbb{E} \min_{U\in\mathcal{C}_d} \frac{1}{p}\sum_{j=1}^p \mathbb{I}\{\widehat{Z}_j U \ne Z_j^*\} \ge c,$$

for some constant $c > 0$.

Our main result in this section shows that the minimax lower bound in Theorem 6.1 can be achieved by an efficient algorithm adaptively over all $\theta^* \in \mathbb{R}^d$ under the signal-to-noise ratio condition $\frac{\Delta_{\min}^2}{d/p+\sqrt{d\log d}} \to \infty$. Specializing Algorithm 1 to the current problem, the iterative procedure is given by the following formula:

$$Z_j^{(t)} = \operatorname*{argmin}_{U\in\mathcal{C}_d} \big\|Y_j - U\widehat{\theta}\big(Z^{(t-1)}\big)\big\|^2, \tag{52}$$

where

$$\widehat{\theta}(Z) = \frac{1}{p}\sum_{j=1}^p Z_j^T Y_j.$$

The computation of (52) is straightforward given that $|\mathcal{C}_d| = d$ and one can simply evaluate $\|Y_j - U\widehat{\theta}(Z^{(t-1)})\|^2$ for each $U \in \mathcal{C}_d$.

6.1. *Conditions*. To analyze the algorithmic convergence, we note that $\mu_j(B^*, U) = \nu_j(B^*, U) = U\theta^*$, $\Delta_j(U, V)^2 = \|(U - V)\theta^*\|^2$ and $\ell(Z, Z^*) = \sum_{j=1}^p \|(Z_j - Z_j^*)\theta^*\|^2$ under the current setting. The error terms that we need to control are

$$F_j(U, V; Z) = \langle \epsilon_j, (U - V)(\widehat{\theta}(Z^*) - \widehat{\theta}(Z))\rangle,$$
$$G_j(U, V; Z) = \langle \widehat{\theta}(Z) - \widehat{\theta}(Z^*), (V^T U - I_d)\theta^*\rangle,$$
$$H_j(U, V) = \langle \widehat{\theta}(Z^*) - \theta^*, (V^T U - I_d)\theta^*\rangle.$$

Here, the noise vector is given by $\epsilon_j = Y_j - Z_j^*\theta^* \sim \mathcal{N}(0, I_d)$. The error terms are controlled by the following lemma.

LEMMA 6.1. *For any $C' > 0$, there exists a constant $C > 0$ only depending on C' such that*

$$\max_{\{Z:\ell(Z,Z^*)\le\tau\}} \sum_{j=1}^p \max_{U\in\mathcal{C}_d\setminus\{Z_j^*\}} \frac{F_j(Z_j^*, U; Z)^2\|\mu_j(B^*, U) - \mu_j(B^*, Z_j^*)\|^2}{\Delta_j(Z_j^*, U)^4\ell(Z, Z^*)}$$
$$\le C\bigg(\frac{(\log d + d/p)\log d}{\Delta_{\min}^4} + \frac{\tau(\log d + d/p)}{p\Delta_{\min}^4}\bigg), \tag{53}$$

$$\max_{\{Z:\ell(Z,Z^*)\le\tau\}} \max_{T\subset[p]} \frac{\tau}{4\Delta_{\min}^2|T| + \tau}$$
$$\times \sum_{j\in T} \max_{U\in\mathcal{C}_d\setminus\{Z_j^*\}} \frac{G_j(Z_j^*, U; Z)^2\|\mu_j(B^*, U) - \mu_j(B^*, Z_j^*)\|^2}{\Delta_j(Z_j^*, U)^4\ell(Z, Z^*)}$$
$$\le C\bigg(\frac{\tau\log d}{p\Delta_{\min}^4} + \frac{\tau^2}{p^2\Delta_{\min}^4}\bigg), \tag{54}$$

and

$$\max_{j\in[p]} \max_{U\in\mathcal{C}_d\setminus\{Z_j^*\}} \frac{|H_j(Z_j^*, U)|}{\Delta_j(Z_j^*, U)^2} \le C\sqrt{\frac{d}{p\Delta_{\min}^2}}, \tag{55}$$

with probability at least $1 - e^{-C'd}$.

From the bounds (53)–(55), we can see that a sufficient condition that Conditions A, B and C hold is $\tau = o(p\Delta_{\min}^2)$ and

$$\frac{\Delta_{\min}^2}{\log d + d/p} \to \infty. \tag{56}$$

In fact, when $d = O(1)$, the above condition is reduced to $\Delta_{\min}^2 \to \infty$, which is the necessary condition for consistency by Theorem 6.1.

Next, we need to bound $\xi_{\text{ideal}}(\delta)$ in Condition D. This is given by the following lemma.

LEMMA 6.2. *Assume $\frac{\Delta_{\min}^2}{\log d + d/p} \to \infty$. Then, for any sequence $\delta_p = o(1)$, we have*

$$\xi_{\text{ideal}}(\xi_p) \le p\exp\bigg(-\big(1 + o(1)\big)\frac{\Delta_{\min}^2}{8}\bigg),$$

with probability at least $1 - \exp(-\Delta_{\min})$.

To summarize, under the signal-to-noise ratio condition (56) and the initialization condition $\tau = o(p\Delta_{\min}^2)$, Conditions A, B, C and D hold simultaneously.

6.2. *Convergence*. With the help of Lemma 6.1 and Lemma 6.2, we can specialize Theorem 3.1 into the following result.

THEOREM 6.2. *Assume* (56) *holds. Suppose* $Z^{(0)}$ *satisfies*

$$\ell\big(Z^{(0)}, Z^*\big) = o\big(p\Delta_{\min}^2\big), \tag{57}$$

with probability at least $1-\eta$. *Then we have*

$$\ell\big(Z^{(t)}, Z^*\big) \le p\exp\bigg(-\big(1+o(1)\big)\frac{\Delta_{\min}^2}{8}\bigg) + \frac{1}{2}\ell\big(Z^{(t-1)}, Z^*\big) \quad \textit{for all } t \ge 1,$$

with probability at least $1-\eta-\exp(-\Delta_{\min})-e^{-d}$.

By the inequality $\frac{1}{p}\sum_{j=1}^p \mathbb{I}\{Z_j \ne Z_j^*\} \le \frac{\ell(Z,Z^*)}{p\Delta_{\min}^2}$, we immediately obtain the following corollary for the Hamming loss.

COROLLARY 6.1. *Assume* (56) *holds. Suppose* $Z^{(0)}$ *satisfies* (57) *with probability at least* $1-\eta$. *Then we have*

$$\frac{1}{p}\sum_{j=1}^p \mathbb{I}\big\{Z_j^{(t)} \ne Z_j^*\big\} \le \exp\bigg(-\big(1+o(1)\big)\frac{\Delta_{\min}^2}{8}\bigg) + 2^{-t} \quad \textit{for all } t \ge 1, \tag{58}$$

with probability at least $1-\eta-\exp(-\Delta_{\min})-e^{-d}$.

According to our lower bound result given by Theorem 6.1, the quantity $\exp(-(1+o(1))\frac{\Delta_{\min}^2}{8})$ is the minimax rate. Moreover, since the loss function $\frac{1}{p}\sum_{j=1}^p \mathbb{I}\{Z_j \ne Z_j^*\}$ takes value in the set $\{j/p : j \in [p] \cup \{0\}\}$, the term 2^{-t} in (58) is negligible as long as $2^{-t} = o(p^{-1})$. We therefore can claim

$$\frac{1}{p}\sum_{j=1}^p \mathbb{I}\big\{Z_j^{(t)} \ne Z_j^*\big\} \le \exp\bigg(-\big(1+o(1)\big)\frac{\Delta_{\min}^2}{8}\bigg) \quad \text{for all } t \ge 3\log p.$$

In other words, the minimax rate is achieved after at most $\lceil 3\log p\rceil$ iterations.

6.3. *Initialization*. To close this section, we discuss a simple initialization procedure that achieves the condition (57). The idea is to find the cyclic shifts for all $j \ge 2$ by using that of $j=1$ as a reference. To be specific, we define $Z_1^{(0)} = I_d$. For each $j \ge 2$, compute

$$Z_j^{(0)} = \operatorname*{argmin}_{U \in \mathcal{C}_d} \big\| Y_1 - Z_j^T Y_j \big\|^2. \tag{59}$$

The estimator (59) has also been discussed by [13]. It is known that (59) does not have optimal statistical error. However, the performance of (59) is sufficient for the purpose of initializing the iterative algorithm (52).

PROPOSITION 6.1. *There exists some* $C > 0$, *such that for any* $\eta > 0$, *we have*

$$\min_{U \in \mathcal{C}_d} \sum_{j=1}^p \big\| (Z_j^{(0)} U - Z_j^*)\theta^* \big\|^2 \le C\frac{p\sqrt{d\log d}}{\eta},$$

with probability at least $1-\eta$.

Proposition 6.1 shows that $Z^{(0)}$ achieves the rate $O(p\sqrt{d\log d})$ for estimating Z^* up to a global cyclic shift $U \in \mathcal{C}_d$. Since $Y_j = Z_j^*\theta^* + \epsilon_j = Z_j^* U^T U\theta^* + \epsilon_j$, the ambiguity of this global cyclic shift cannot be avoided.

In order that the initialization condition (57) is satisfied, we shall consider the signal-to-noise ratio condition

$$\frac{\Delta_{\min}^2}{\sqrt{d\log d} + d/p} \to \infty. \tag{60}$$

Note that (60) implies (56) and thus the algorithmic convergence holds. Given the condition (60), we can take $\eta = \frac{\sqrt{d\log d}}{\Delta_{\min}^2}$ in Proposition 6.1. Then, under (60), we have

$$\min_{U\in\mathcal{C}_d} \sum_{j=1}^{p} \|(Z_j^{(0)}U - Z_j^*)\theta^*\|^2 = o(p\Delta_{\min}^2),$$

with probability at least $1 - \frac{\sqrt{d\log d}}{\Delta_{\min}^2}$. This means there exists some $U \in \mathcal{C}_d$, such that the initial estimator $\{Z_j^{(0)}U\}$ recovers $\{Z_j^*\}$ after a global shift with an error that satisfies the condition (57). Therefore, Corollary 6.1 implies that $\min_{U\in\mathcal{C}_d} \frac{1}{p}\sum_{j=1}^{p} \mathbb{I}\{Z_j^{(t)}U \neq Z_j^*\}$ converges to the minimax error with a linear rate under the signal-to-noise ratio condition (60).

7. Group synchronization. In this section, we study a general class of problems called group synchronization. Given a group $(\mathcal{G}, \circ)$ and group elements $g_1, \ldots, g_p \in \mathcal{G}$, we observe noisy versions of $g_i \circ g_j^{-1}$, and the goal is to recover the group elements $g_1, \ldots, g_p$. It turns out our general framework is particularly suitable to solve group synchronization, at least for discrete groups. We will consider the following three representative examples:

1. $\mathbb{Z}_2$ synchronization. This is the simplest example of group synchronization, and it is closely related to the more general phase/angular synchronization problem [12]. The group only consists of two elements $\{-1, 1\}$, and the group operation is the ordinary product.

2. $\mathbb{Z}/k\mathbb{Z}$ synchronization. Also known as joint alignment from pairwise differences, $\mathbb{Z}/k\mathbb{Z}$ synchronization was first considered by [23]. The group consists of elements $\{0, 1, 2, \ldots, k-1\}$ with group operation $g \circ h = g + h (\mathrm{mod}\, k)$.

3. Permutation synchronization. As one of the most popular methods for multiple image alignment, permutation synchronization was first proposed by [73]. In this example, the group contains all permutations of $[d]$, and the group operation is the composition of two permutations.

Given its importance in applied mathematics and engineering, group synchronization has been extensively studied in the literature [3, 4, 12, 24, 35, 56, 58, 59, 76, 89, 95]. Most approaches in the literature are based on semidefinite programming (SDP) and other forms of convex relaxations. In terms of statistical guarantees, the literature is mainly focused on conditions of exact recovery. In fact, for the three examples that we list above, the minimax rates are unknown with $\mathbb{Z}_2$ synchronization being the only exception. In this section, we will show Algorithm 1 can be specialized to the three models and can achieve the minimax rate of each one. Even for $\mathbb{Z}_2$ synchronization, our result offers some new insight of the problem. The minimax rate of $\mathbb{Z}_2$ synchronization is achieved by an SDP procedure [35] in the literature. In comparison, Algorithm 1 leads to a much simpler power method, which is easier to implement in practice.

There are several different options of noise models in the literature. The most standard and popular choice is $Y_{ij} = g_i \circ g_j^{-1} + \sigma W_{ij}$ with W_{ij} being a Gaussian element. However, it is

also natural to restrict the noisy observation of $g_i \circ g_j^{-1}$ to be an element of the group $\mathcal{G}$. One way to achieve this is the noise model [80],

$$Y_{ij} = \begin{cases} g_i \circ g_j^{-1} & \text{with probability } q, \\ \text{Uniform}(\mathcal{G}) & \text{with probability } 1-q. \end{cases} \tag{61}$$

Another way is through projection [4]. Namely, $Y_{ij} = \mathcal{P}_{\mathcal{G}}(g_i \circ g_j^{-1} + \sigma W_{ij})$, where $\mathcal{P}_{\mathcal{G}}$ is a projection onto $\mathcal{G}$ with respect to the ℓ_2 norm. In addition, one can also consider partial observations on a random graph [24]. As argued in [4], these noise models are all equivalent to

$$Y_{ij} = \lambda g_i \circ g_j^{-1} + W_{ij}, \tag{62}$$

with some $\lambda \in \mathbb{R}$ depending on σ or q and some additive noise W_{ij} that is sub-Gaussian.[2] For simplicity, we thus consider the noise model (62) with W_{ij} being a standard Gaussian element. The results we obtain in this section can all be extended with a sub-Gaussian W_{ij} to include more general noise settings.

In the rest of this section, we focus on the setting of $\mathbb{Z}_2$ synchronization. The results of $\mathbb{Z}/k\mathbb{Z}$ synchronization and permutation synchronization will be given in the Supplementary Material [46] due to page limit. Consider the observations $Y_{ij} \sim \mathcal{N}(\lambda^* z_i^* z_j^*, 1)$ independently for all $1 \le i < j \le p$ with $z_i^* \in \{-1, 1\}$ and $\lambda^* \in \mathbb{R}$. Using matrix notation, we can write $Y = \lambda^* z^* z^{*T} + W$, where W is a symmetric matrix such that $W_{ij} = W_{ji} \sim \mathcal{N}(0,1)$ for all $1 \le i < j \le p$ and $W_{ii} = 0$ for all $i \in [p]$. This is the simplest group synchronization problem, and is closely related to the problem of angular synchronization [12]. The minimax lower bound of this problem has been recently obtained by [35].

THEOREM 7.1 (Fei and Chen [35]). *If $p\lambda^{*2} \to \infty$, we have*

$$\inf_{\widehat{z}} \sup_{z^*} \mathbb{E}\left(\frac{1}{p} \sum_{j=1}^{p} \mathbb{I}\{\widehat{z}_j \ne z_j^*\} \wedge \frac{1}{p} \sum_{j=1}^{p} \mathbb{I}\{\widehat{z}_j \ne -z_j^*\} \right) \ge \exp\left(-\frac{(1+o(1))p\lambda^{*2}}{2} \right).$$

Otherwise, if $p\lambda^{*2} = O(1)$, we then have

$$\inf_{\widehat{z}} \sup_{z^*} \mathbb{E}\left(\frac{1}{p} \sum_{j=1}^{p} \mathbb{I}\{\widehat{z}_j \ne z_j^*\} \wedge \frac{1}{p} \sum_{j=1}^{p} \mathbb{I}\{\widehat{z}_j \ne -z_j^*\} \right) \ge c,$$

for some constant $c > 0$.

The result is coherent with the necessary condition of weak recovery (i.e., to find a $\widehat{z}$ that is correlated with z^*) $p\lambda^{*2} \to \infty$ and strong/exact recover (i.e., to find a $\widehat{z}$ that equals z^* up to a sign) $p\lambda^{*2} > 2\log p$ in the literature [12, 76]. It was proved by [35] that the minimax rate can be achieved by a semidefinite programming (SDP). In this section, we show that a simpler power iteration method, a special case of our general iterative algorithm, also achieves this minimax rate.

As is discussed in Section 2, the $\mathbb{Z}_2$ synchronization model can be equivalently represented as $Y = z^*(\beta^*)^T + W$ with $\beta^* = \lambda^* z^* \in \mathcal{B}_{z^*} = \{\beta = \lambda z^* : \lambda \in \mathbb{R}\}$. The resulting iterative algorithm can be summarized as

$$z_j^{(t)} = \operatorname*{argmin}_{a \in \{-1,1\}} \left\| Y_j - a\widehat{\beta}(z^{(t-1)}) \right\|^2, \tag{63}$$

[2]The equivalence between the projection noise and (62) is only true for some special groups

where for any $z \in \{-1, 1\}^p$, we use the notation

$$\widehat{\beta}(z) = \frac{z^T Y z}{p^2} z.$$

The iterative procedure (63) is equivalent to the power method (20). The power method enjoys good theoretical properties in the setting of angular synchronization [94], but whether it can achieve the minimax rate of $\mathbb{Z}_2$ synchronization is unknown in the literature.

7.1. *Conditions*. To analyze the algorithmic convergence of (63), we note that $\|\beta^*\|^2 = |\lambda^*|^2 p$. Then $\mu_j(B^*, a) = \nu_j(B^*, a) = a\beta^*$, $\Delta_j(a, b)^2 = (a - b)^2 \|\beta^*\|^2$, $\Delta_{\min}^2 = \min_{a \neq b} \Delta_j(a, b)^2 = 4\|\beta^*\|^2$, and

$$\ell(z, z^*) = \sum_{j=1}^{p} (z_j - z_j^*)^2 \|\beta^*\|^2 = p|\lambda^*|^2 \sum_{j=1}^{p} (z_j - z_j^*)^2, \tag{64}$$

under the current setting. The error terms that we need to control are given by the formulas (26)–(28). Here, the noise vector is given by $\epsilon_j = W_j$, the jth column of the error matrix W. The error terms are controlled by the following lemma.

LEMMA 7.1. *For any $C' > 0$, there exists a constant $C > 0$ only depending on C' such that*

$$\begin{aligned} &\max_{\{z:\ell(z,z^*)\le\tau\}} \sum_{j=1}^{p} \max_{b\in\{-1,1\}\setminus\{z_j^*\}} \frac{F_j(z_j^*, b; z)^2 \|\mu_j(B^*, b) - \mu_j(B^*, z_j^*)\|^2}{\Delta_j(z_j^*, b)^4 \ell(z, z^*)} \\ &\quad \le C\biggl(\frac{1}{p\lambda^{*2}} + \frac{1}{p^2\lambda^{*4}}\biggr), \end{aligned} \tag{65}$$

$$\begin{aligned} &\max_{\{z:\ell(z,z^*)\le\tau\}} \max_{T\subset[p]} \frac{\tau}{4\Delta_{\min}^2|T| + \tau} \\ &\qquad \times \sum_{j\in T} \max_{b\in\{-1,1\}\setminus\{z_j^*\}} \frac{G_j(z_j^*, b; z)^2 \|\mu_j(B^*, b) - \mu_j(B^*, z_j^*)\|^2}{\Delta_j(z_j^*, b)^4 \ell(z, z^*)} \\ &\quad \le C\biggl(\frac{\tau}{\lambda^{*2} p^2} + \frac{\tau}{\lambda^{*4} p^3}\biggr), \end{aligned} \tag{66}$$

and

$$\max_{j\in[p]} \max_{b\neq z_j^*} \frac{|H_j(z_j^*, b)|}{\Delta_j(z_j^*, b)^2} \le C \frac{1}{\sqrt{p\lambda^{*2}}}, \tag{67}$$

with probability at least $1 - e^{-C'p}$.

From the bounds (65)–(67), we can see that a sufficient condition that Conditions A, B and C hold is $p\lambda^{*2} \to \infty$ and $\tau = o(p^2\lambda^{*2})$. Note that $p\lambda^{*2} \to \infty$ is also the necessary condition for consistency according to Theorem 7.1.

Next, we need to bound $\xi_{\text{ideal}}(\delta)$ in Condition D. This is given by the following lemma.

LEMMA 7.2. *Assume $p\lambda^{*2} \to \infty$. Then, for any sequence $\delta_p = o(1)$, we have*

$$\xi_{\text{ideal}}(\delta_p) \le p \exp\biggl(-\bigl(1 + o(1)\bigr)\frac{p\lambda^{*2}}{2}\biggr),$$

with probability at least $1 - \exp(-\sqrt{p\lambda^{*2}})$.

Again, the same condition $p\lambda^{*2} \to \infty$ is required for Lemma 7.2. Thus, under the condition $p\lambda^{*2} \to \infty$, Conditions A, B, C and D hold simultaneously.

7.2. *Convergence.* With the help of Lemma 7.1 and Lemma 7.2, we can specialize Theorem 3.1 into the following result.

THEOREM 7.2. *Assume* $p\lambda^{*2} \to \infty$. *Suppose* $z^{(0)}$ *satsifies*

$$\ell\big(z^{(0)}, z^*\big) = o\big(p^2\lambda^{*2}\big), \tag{68}$$

with probability at least $1 - \eta$. *Then we have*

$$\ell\big(z^{(t)}, z^*\big) \le p\exp\bigg(-\big(1 + o(1)\big)\frac{p\lambda^{*2}}{2}\bigg) + \frac{1}{2}\ell\big(z^{(t-1)}, z^*\big) \quad \textit{for all } t \ge 1,$$

with probability at least $1 - \eta - \exp(-\sqrt{p\lambda^{*2}}) - e^{-p}$.

By the inequality, $\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{z_j \ne z_j^*\} \le \frac{\ell(z,z^*)}{p^2\lambda^{*2}}$, we immediately obtain the following corollary for the Hamming loss.

COROLLARY 7.1. *Assume* $p\lambda^{*2} \to \infty$. *Suppose* $z^{(0)}$ *satisfies* (68) *with probability at least* $1 - \eta$. *Then*

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\big\{z_j^{(t)} \ne z_j^*\big\} \le \exp\bigg(-\big(1 + o(1)\big)\frac{p\lambda^{*2}}{2}\bigg) + 2^{-t} \quad \textit{for all } t \ge 1, \tag{69}$$

with probability at least $1 - \eta - \exp(-\sqrt{p\lambda^{*2}}) - e^{-p}$.

By the property of the Hamming loss, the algorithmic error 2^{-t} is negligible after $\lceil 3\log p\rceil$ iterations, and we have

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\big\{z_j^{(t)} \ne z_j^*\big\} \le \exp\bigg(-\big(1 + o(1)\big)\frac{p\lambda^{*2}}{2}\bigg) \quad \text{for all } t \ge 3\log p.$$

Thus, the minimax rate is achieved given that the initialization condition (68) is satisfied.

7.3. *Initialization.* Observe that the expectation of Y has a rank one structure. Thus, a natural initialization procedure is to extract the information of z^* by computing the leading eigenvector of Y. Let $\widehat{u} = \operatorname{argmax}_{\|u\|=1} u^T Y u$, and we define $z_j^{(0)} = \operatorname{sign}(\widehat{u}_j)$ for all $j \in [p]$. The behavior of $\widehat{z}^{(0)}$ has been analyzed by [3] when $p\lambda^{*2} > 2\log p$. Without this condition, we show $\widehat{z}^{(0)}$ can be used as a good initialization for the iterative algorithm.

PROPOSITION 7.1. *For any* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* C' *such that*

$$\ell\big(z^{(0)}, z^*\big) \wedge \ell\big(z^{(0)}, -z^*\big) \le Cp,$$

with probability at least $1 - e^{-C'p}$.

Proposition 7.1 shows that $z^{(0)}$ achieves the rate $O(p)$ for estimating z^* up to a change of sign. Interestingly, the initialization condition (68) is satisfied as long as $p\lambda^{*2} \to \infty$, the same condition that we use for both the lower bound (Theorem 7.1) and the algorithmic convergence (7.2). Therefore, by Corollary 7.1, $\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{z_j^{(t)} \ne z_j^*\} \wedge \frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{z_j^{(t)} \ne -z_j^*\}$ converges to the minimax rate with a linear rate under the condition $p\lambda^{*2} \to \infty$.

8. Discussion. In this paper, we show that a number of different discrete structure recovery problems can be unified into a single framework, and a general iterative algorithm is proved to achieve the optimal statistical error rate for each problem. In addition to all the examples covered by the paper, we expect our framework will lead to applications in many other statistical models, thanks to the flexibility of the general structured linear models (4). However, it is also worthwhile to note some important limitations of our proposed framework in the end of the paper. We compile four major points, listed below:

1. *Parameter estimation*. While our iterative algorithm is designed for estimating the discrete structure z^*, it also outputs an estimator for the continuous model parameter. A natural question is whether this estimator enjoys any statistical optimality for estimating B^*. The answer to this question is a very clear no. As a concrete example, it is known that the k-means algorithm leads to suboptimal parameter estimation due to the bias resulted from the clustering error [62]. Instead, for parameter estimation, one should use the EM algorithm for the Gaussian mixture model [33, 86]. This phenomenon is also known as the Neyman–Scott paradox in linear mixed models. Basically, for optimal global parameter estimation, one should integrate out the local latent variables instead of optimizing over them. In general, the iterative algorithm proposed in the paper is only statistically optimal for recovering the discrete structure z^*.

2. *Models with extremely weak SNR*. The examples that we analyze in the paper all exhibit reasonable signal-to-noise ratio behaviors. There are other examples that fit perfectly in the framework of structured linear models, but do not lead to convergent iterative algorithms. Consider a shuffled regression problem with independent observations $x_i \sim \mathcal{N}(0, I_d)$ and $y_i | x_{z_i^*} \sim \mathcal{N}(x_{z_i^*}^T \beta^*, 1)$ for $i = 1, \ldots, n$. In this model, the vector z^* is a label permutation that links the covariates to the response. To recover z^*, the quantity $\|\beta^*\|^2$ serves as the signal strength of the problem. It was proved by [74] that the recovery of z^* is only possible under the signal-to-noise ratio condition $\|\beta^*\|^2 \geq e^{O(n)}$. The problem has such a weak signal strength, and our analysis of the error terms $F_j(a, b; z)$, $G_j(a, b; z)$ and $H_j(a, b)$ simply breaks down. In fact, statistical recovery of z^* in polynomial time under the condition $\|\beta^*\|^2 \geq e^{O(n)}$ still remains an open problem in the literature.

3. *Models with nonlocal discrete structure*. Consider a change-point problem $Y \sim \mathcal{N}(\theta^*, I_n)$ with θ^* having a piecewise constant structure. In other words, there exists $z_2^*, \ldots, z_n^* \in \{0, 1\}$ such that $\theta_i^* = \theta_{i-1}^*$ for all $z_i^* = 1$. This model can be easily written as a structured linear model, but there is no suitable iterative algorithm to recover the change-point structure encoded in $z_2^*, \ldots, z_n^*$. The reason is the lack of local statistic T_i that is sufficient for z_i^*. The change-point structure is a discrete but nonlocal structure. The amount of information for z_i^* is dependent on the locations of the previous and the next change points, which are further determined by other z_i's. Our iterative algorithm is not suitable for recovering such nonlocal discrete structures. An appropriate algorithm for this problem is dynamic programming [39].

4. *Link function*. Though the Gaussianity assumption in the paper can all be relaxed to sub-Gaussian errors, we still require both the structured linear model (4) and the local statistic (5) to have additive error structures. This requirement is coherent with the iterative algorithm, since both iteration steps (8) and (9) are least-squares optimization. Problems such as variable selection in generalized linear models and ranking in Bradley–Terry–Luce model [19, 63] involve link functions that are not identity. This poses new challenges in the error analysis in addition to the modification of the iterative algorithm.

While some of the listed points may be addressed by appropriate extensions of our framework, others may require a fundamentally different approach to the problem. We hope the above discussion not only highlights the critical features of our proposed framework, but also leads to potential future research projects in discrete structure recovery.

Funding. Research of Chao Gao is supported in part by NSF CAREER award DMS-1847590 and NSF grant CCF-1934931. Research of Anderson Y. Zhang is supported in part by NSF grant DMS-2112988.

SUPPLEMENTARY MATERIAL

Supplement to "Iterative algorithm for discrete structure recovery" (DOI: 10.1214/21-AOS2140SUPP; .pdf). The supplement [46] includes a few more examples and all the technical proofs. We first analyze approximate ranking in Appendix A. $\mathbb{Z}/k\mathbb{Z}$ synchronization and permutation synchronization are studied in Appendix B and Appendix C, respectively. We then prove Theorem 3.1 in Appendix D. The rest of the proofs are organized from Appendix E to Appendix I.

REFERENCES

[1] ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

[2] ABBE, E., BENDORY, T., LEEB, W., PEREIRA, J. M., SHARON, N. and SINGER, A. (2019). Multireference alignment is easier with an aperiodic translation distribution. *IEEE Trans. Inf. Theory* **65** 3565–3584. MR3959006 https://doi.org/10.1109/TIT.2018.2889674

[3] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474. MR4124330 https://doi.org/10.1214/19-AOS1854

[4] ABBE, E., MASSOULIÉ, L., MONTANARI, A., SLY, A. and SRIVASTAVA, N. (2018). Group synchronization on grids. *Math. Stat. Learn.* **1** 227–256. MR4059722 https://doi.org/10.4171/msl/6

[5] AERON, S., SALIGRAMA, V. and ZHAO, M. (2010). Information theoretic bounds for compressed sensing. *IEEE Trans. Inf. Theory* **56** 5111–5130. MR2808668 https://doi.org/10.1109/TIT.2010.2059891

[6] AGUERREBERE, C., DELBRACIO, M., BARTESAGHI, A. and SAPIRO, G. (2016). Fundamental limits in multi-image alignment. *IEEE Trans. Signal Process.* **64** 5707–5722. MR3548763 https://doi.org/10.1109/TSP.2016.2600517

[7] ALOISE, D., DESHPANDE, A., HANSEN, P. and POPAT, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Mach. Learn.* **75** 245–248.

[8] ARTHUR, D. and VASSILVITSKII, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

[9] AWASTHI, P. and SHEFFET, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Lecture Notes in Computer Science* **7408** 37–49. Springer, Heidelberg. MR3003539 https://doi.org/10.1007/978-3-642-32512-0_4

[10] BAGARIA, V., DING, J., TSE, D., WU, Y. and XU, J. (2020). Hidden Hamiltonian cycle recovery via linear programming. *Oper. Res.* **68** 53–70. MR4059492 https://doi.org/10.1287/opre.2019.1886

[11] BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

[12] BANDEIRA, A. S., BOUMAL, N. and SINGER, A. (2017). Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Math. Program.* **163** 145–167. MR3632977 https://doi.org/10.1007/s10107-016-1059-6

[13] BANDEIRA, A. S., CHARIKAR, M., SINGER, A. and ZHU, A. (2014). Multireference alignment using semidefinite programming. In *ITCS'*14*—Proceedings of the* 2014 *Conference on Innovations in Theoretical Computer Science* 459–470. ACM, New York. MR3359498

[14] BANDEIRA, A. S., NILES-WEED, J. and RIGOLLET, P. (2019). Optimal rates of estimation for multireference alignment. *Math. Stat. Learn.* **2** 25–75. MR4073147 https://doi.org/10.4171/msl/11

[15] BELLEC, P. C., LECUÉ, G. and TSYBAKOV, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642. MR3852663 https://doi.org/10.1214/17-AOS1670

[16] BENDORY, T., BOUMAL, N., MA, C., ZHAO, Z. and SINGER, A. (2018). Bispectrum inversion with application to multireference alignment. *IEEE Trans. Signal Process.* **66** 1037–1050. MR3771661 https://doi.org/10.1109/TSP.2017.2775591

[17] BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27** 265–274. MR2559726 https://doi.org/10.1016/j.acha.2009.04.002

[18] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717 https://doi.org/10.1214/15-AOAS842

[19] BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925 https://doi.org/10.2307/2334029

[20] BRAVERMAN, M. and MOSSEL, E. (2008). Noisy sorting without resampling. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 268–276. ACM, New York. MR2485312

[21] BRODER, A. Z., FRIEZE, A. M. and SHAMIR, E. (1994). Finding hidden Hamiltonian cycles. *Random Structures Algorithms* **5** 395–410. MR1277610 https://doi.org/10.1002/rsa.3240050303

[22] BUTUCEA, C., NDAOUD, M., STEPANOVA, N. A. and TSYBAKOV, A. B. (2018). Variable selection with Hamming loss. *Ann. Statist.* **46** 1837–1875. MR3845003 https://doi.org/10.1214/17-AOS1572

[23] CHEN, Y. and CANDÈS, E. J. (2018). The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Comm. Pure Appl. Math.* **71** 1648–1714. MR3847751 https://doi.org/10.1002/cpa.21760

[24] CHEN, Y., GUIBAS, L. J. and HUANG, Q.-X. (2014). Near-optimal joint object matching via convex relaxation. Preprint. Available at arXiv:1402.1473.

[25] COLLIER, O. and DALALYAN, A. S. (2016). Minimax rates in permutation estimation for feature matching. *J. Mach. Learn. Res.* **17** Paper No. 6, 31 pp. MR3482926

[26] CONTE, D., FOGGIA, P., SANSONE, C. and VENTO, M. (2004). Thirty years of graph matching in pattern recognition. *Int. J. Pattern Recognit. Artif. Intell.* **18** 265–298.

[27] DASGUPTA, S. (2008). *The Hardness of k-Means Clustering*. Department of Computer Science and Engineering, Univ. California.

[28] DASKALAKIS, C., TZAMOS, C. and ZAMPETAKIS, M. (2016). Ten steps of EM suffice for mixtures of two Gaussians. Preprint. Available at arXiv:1609.00368.

[29] DAWID, A. P. and SKENE, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **28** 20–28.

[30] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

[31] DERUMIGNY, A. (2018). Improved bounds for square-root Lasso and square-root slope. *Electron. J. Stat.* **12** 741–766. MR3769194 https://doi.org/10.1214/18-EJS1410

[32] DING, J., MA, Z., WU, Y. and XU, J. (2021). Efficient random graph matching via degree profiles. *Probab. Theory Related Fields* **179** 29–115. MR4221654 https://doi.org/10.1007/s00440-020-00997-4

[33] DWIVEDI, R., HO, N., KHAMARU, K., WAINWRIGHT, M. J., JORDAN, M. I. and YU, B. (2020). Singularity, misspecification and the convergence rate of EM. *Ann. Statist.* **48** 3161–3182. MR4185804 https://doi.org/10.1214/19-AOS1924

[34] FEI, Y. and CHEN, Y. (2019). Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Trans. Inf. Theory* **65** 551–571. MR3901009 https://doi.org/10.1109/TIT.2018.2839677

[35] FEI, Y. and CHEN, Y. (2020). Achieving the Bayes error rate in synchronization and block models by SDP, robustly. *IEEE Trans. Inf. Theory* **66** 3929–3953. MR4115142 https://doi.org/10.1109/TIT.2020.2966438

[36] FLETCHER, A. K., RANGAN, S. and GOYAL, V. K. (2009). Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inf. Theory* **55** 5758–5772. MR2597192 https://doi.org/10.1109/TIT.2009.2032726

[37] FOUCART, S. (2011). Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM J. Numer. Anal.* **49** 2543–2563. MR2873246 https://doi.org/10.1137/100806278

[38] FRANK, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*: *Visualization of Biological Molecules in Their Native State*. Oxford Univ. Press, London.

[39] FRIEDRICH, F., KEMPE, A., LIEBSCHER, V. and WINKLER, G. (2008). Complexity penalized M-estimation: Fast computation. *J. Comput. Graph. Statist.* **17** 201–224. MR2424802 https://doi.org/10.1198/106186008X285591

[40] GAO, C. (2017). Phase transitions in approximate ranking. Preprint. Available at arXiv:1711.11189.

[41] GAO, C., LU, Y. and ZHOU, D. (2016). Exact exponent in optimal rates for crowdsourcing. In *International Conference on Machine Learning* 603–611.

[42] GAO, C. and MA, Z. (2021). Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statist. Sci.* **36** 16–33. MR4194201 https://doi.org/10.1214/19-STS736

[43] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45 pp. MR3687603

[44] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2018). Community detection in degree-corrected block models. *Ann. Statist.* **46** 2153–2185. MR3845014 https://doi.org/10.1214/17-AOS1615

[45] GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2020). A general framework for Bayes structured linear models. *Ann. Statist.* **48** 2848–2878. MR4152123 https://doi.org/10.1214/19-AOS1909
[46] GAO, C. and ZHANG, A. Y. (2022). Supplement to "Iterative algorithm for discrete structure recovery." https://doi.org/10.1214/21-AOS2140SUPP
[47] GIRAUD, C. and VERZELEN, N. (2018). Partial recovery bounds for clustering with the relaxed K-means. *Math. Stat. Learn.* **1** 317–374. MR4059724
[48] GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826. MR1908073 https://doi.org/10.1073/pnas.122653799
[49] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inf. Theory* **62** 2788–2797. MR3493879 https://doi.org/10.1109/TIT.2016.2546280
[50] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. Inf. Theory* **62** 5918–5937. MR3552431 https://doi.org/10.1109/TIT.2016.2594812
[51] HARTIGAN, J. A. (1975). *Clustering Algorithms*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0405726
[52] JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180 https://doi.org/10.1214/11-AOS947
[53] KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., PIATKO, C. D., SILVERMAN, R. and WU, A. Y. (2004). A local search approximation algorithm for k-means clustering. *Comput. Geom.* **28** 89–112. MR2062789 https://doi.org/10.1016/j.comgeo.2004.03.003
[54] KUMAR, A. and KANNAN, R. (2010). Clustering with spectral norm and the k-means algorithm. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 299–308. IEEE Computer Soc., Los Alamitos, CA. MR3025203
[55] KUMAR, A., SABHARWAL, Y. and SEN, S. (2004). A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Annual Symposium on Foundations of Computer Science* **45** 454–462. IEEE Computer Society Press.
[56] LERMAN, G. and SHI, Y. (2019). Robust group synchronization via cycle-edge message passing. Preprint. Available at arXiv:1912.11347.
[57] LESKOVEC, J., LANG, K. J. and MAHONEY, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the* 19*th International Conference on World Wide Web* 631–640. ACM, New York.
[58] LING, S. (2020). Solving orthogonal group synchronization via convex and low-rank optimization: Tightness and landscape analysis. Preprint. Available at arXiv:2006.00902.
[59] LING, S. (2020). Near-optimal performance bounds for orthogonal and permutation group synchronization via spectral methods. Preprint. Available at arXiv:2008.05341.
[60] LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28** 129–137. MR0651807 https://doi.org/10.1109/TIT.1982.1056489
[61] LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087 https://doi.org/10.1214/08-EJS177
[62] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. Preprint. Available at arXiv:1612.02099.
[63] LUCE, R. D. (2012). *Individual Choice Behavior*: *A Theoretical Analysis*. Wiley, New York. MR0108411
[64] MAHAJAN, M., NIMBHORKAR, P. and VARADARAJAN, K. (2012). The planar k-means problem is NP-hard. *Theoret. Comput. Sci.* **442** 13–21. MR2927097 https://doi.org/10.1016/j.tcs.2010.05.034
[65] MAO, C., WEED, J. and RIGOLLET, P. (2018). Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory* 2018. *Proc. Mach. Learn. Res.* (*PMLR*) **83** 27. Proceedings of Machine Learning Research PMLR. MR3857331
[66] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281
[67] MONTANARI, A. and SEN, S. (2016). Semidefinite programs on sparse random graphs and their application to community detection. In *STOC'*16*—Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 814–827. ACM, New York. MR3536616 https://doi.org/10.1145/2897518.2897548
[68] MONTEILLER, P., CLAICI, S., CHIEN, E., MIRZAZADEH, F., SOLOMON, J. M. and YUROCHKIN, M. (2019). Alleviating label switching with optimal transport. In *Advances in Neural Information Processing Systems* 13634–13644.
[69] MOSSEL, E., NEEMAN, J. and SLY, A. (2014). Consistency thresholds for binary symmetric block models. Preprint. Available at arXiv:1407.1591.
[70] MOSSEL, E., NEEMAN, J. and SLY, A. (2018). A proof of the block model threshold conjecture. *Combinatorica* **38** 665–708. MR3876880 https://doi.org/10.1007/s00493-016-3238-8

[71] NDAOUD, M. (2018). Sharp optimal recovery in the two Gaussian mixture model. Preprint. Available at arXiv:1812.08078.

[72] NDAOUD, M. and TSYBAKOV, A. B. (2020). Optimal variable selection and adaptive noisy compressed sensing. *IEEE Trans. Inf. Theory* **66** 2517–2532. MR4087700 https://doi.org/10.1109/TIT.2020.2965738

[73] PACHAURI, D., KONDOR, R. and SINGH, V. (2013). Solving the multi-way matching problem by permutation synchronization. In *Advances in Neural Information Processing Systems* 1860–1868.

[74] PANANJADY, A., WAINWRIGHT, M. J. and COURTADE, T. A. (2018). Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Trans. Inf. Theory* **64** 3286–3300. MR3798377 https://doi.org/10.1109/TIT.2017.2776217

[75] PERRY, A., WEED, J., BANDEIRA, A. S., RIGOLLET, P. and SINGER, A. (2019). The sample complexity of multireference alignment. *SIAM J. Math. Data Sci.* **1** 497–517. MR4002723 https://doi.org/10.1137/18M1214317

[76] PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2016). Optimality and sub-optimality of pca for spiked random matrices and synchronization. Preprint. Available at arXiv:1609.05573.

[77] RAD, K. R. (2011). Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Trans. Inf. Theory* **57** 4672–4679. MR2840483 https://doi.org/10.1109/TIT.2011.2145670

[78] SALIGRAMA, V. and ZHAO, M. (2011). Thresholded basis pursuit: LP algorithm for order-wise optimal support recovery for sparse and approximately sparse signals from noisy random measurements. *IEEE Trans. Inf. Theory* **57** 1567–1586. MR2815835 https://doi.org/10.1109/TIT.2011.2104512

[79] SIGWORTH, F. J. (1998). A maximum-likelihood approach to single-particle image refinement. *J. Struct. Biol.* **122** 328–339.

[80] SINGER, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30** 20–36. MR2737931 https://doi.org/10.1016/j.acha.2010.02.001

[81] SU, W. and CANDÈS, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953 https://doi.org/10.1214/15-AOS1397

[82] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory* **55** 5728–5741. MR2597190 https://doi.org/10.1109/TIT.2009.2032816

[83] WANG, W., WAINWRIGHT, M. J. and RAMCHANDRAN, K. (2010). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theory* **56** 2967–2979. MR2683451 https://doi.org/10.1109/TIT.2010.2046199

[84] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689 https://doi.org/10.1214/08-AOS646

[85] WU, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867 https://doi.org/10.1214/aos/1176346060

[86] WU, Y. and ZHOU, H. H. (2019). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $\sqrt{n}$ iterations. Preprint. Available at arXiv:1908.10935.

[87] XU, J., HSU, D. J. and MALEKI, A. (2016). Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems* 2676–2684.

[88] XU, J., HSU, D. J. and MALEKI, A. (2018). Benefits of over-parameterization with EM. In *Advances in Neural Information Processing Systems* 10662–10672.

[89] YAN, J., CHO, M., ZHA, H., YANG, X. and CHU, S. M. (2015). Multi-graph matching via affinity optimization with graduated consistency regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **38** 1228–1242.

[90] YUN, S.-Y. and PROUTIERE, A. (2014). Accurate community detection in the stochastic block model via spectral algorithms. Preprint. Available at arXiv:1412.7335.

[91] ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428

[92] ZHANG, A. Y. and ZHOU, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *Ann. Statist.* **48** 2575–2598. MR4152113 https://doi.org/10.1214/19-AOS1898

[93] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[94] ZHONG, Y. and BOUMAL, N. (2018). Near-optimal bounds for phase synchronization. *SIAM J. Optim.* **28** 989–1016. MR3782406 https://doi.org/10.1137/17M1122025

[95] ZHOU, X., ZHU, M. and DANIILIDIS, K. (2015). Multi-image matching via fast alternating minimization. In *Proceedings of the IEEE International Conference on Computer Vision* 4032–4040.