Information-theoretic Limits of Subspace Clustering

Kwangjun Ahn
Dept. of Mathematical Sciences
KAIST
kjahnkorea@kaist.ac.kr

Kangwook Lee
School of EE
KAIST
kw1jjang@kaist.ac.kr

Changho Suh
School of EE
KAIST
chsuh@kaist.ac.kr

***Abstract*—Subspace clustering is a celebrated problem that comes up in a variety of applications such as motion segmentation and face clustering. The goal of the problem is to find clusters in different subspaces from similarity measurements across data points. While the algorithmic aspect of this problem has been extensively studied in the literature, the information-theoretic limit on the number of similarities required for reliable clustering has been unknown. In this paper, we translate the problem into an instance of community recovery in hypergraphs, and characterize the sharp threshold on the limit required for exact subspace clustering. Moreover, we present a computationally efficient algorithm that achieves the fundamental limit.**

I. Introduction

Subspace clustering is a popular problem of which the task is to cluster n mixed data points that approximately lie in a union of affine subspaces by measuring similarities between data points. Subspace clustering has been applied to a variety of applications such as motion segmentation [1] and face clustering [2], where data points corresponding to the same class (tracked points on a moving object or faces of a person) lie on a single subspace. See [3] and references therein for a thorough survey of applications.

A common procedure in the existing algorithms for subspace clustering [4]–[7] includes construction of the affinity tensor of the d-th order ($d \geq 2$) whose entries represent similarities between every d data points. Since this construction suffers from a high computational complexity that scales like n^d, the prior algorithms face a challenge in applications with large data sets.

To address such computational challenge, sampling-based approaches, which construct the affinity tensor with sampled entries, are proposed in [4], [8]. While those approaches successfully reduce the computational complexity in practice, no theoretical guarantees are provided. Recently, Ghoshdastidar and Dukkipati [9] demonstrate that sampling $\Omega(n \log^2 n)$ entries suffices for clustering up to a vanishing fraction of wrongly clustered data points; however, whether or not the provable sample complexity is optimal has been unknown. Further, sample complexity conditions for *exact* clustering have been unexplored.

In this work, we translate the subspace clustering problem into an instance of the problem of community recovery in hypergraphs, and characterize the fundamental limit on sample complexity required for *exact* subspace clustering. We show that such fundamental limit reads $\frac{2^{d-2}}{d} \frac{n \log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}$, where θ is a noise-related parameter in similarity measurements (to be detailed later; see (1) in particular). We also present a polynomial-time algorithm that achieves the limit, thereby demonstrating no computational barrier.

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (2017-0-00694, Coding for High-Speed Distributed Networks).

A. Connection to Community Recovery

Community recovery problems seek to find hidden communities given a graph and have been extensively studied [10]–[13]. Recently, [14] investigates a hypergraph setting in which edges cover more than two nodes. Drawing parallels between data points and nodes in a hypergraph, we translate subspace clustering into an instance of community recovery in hypergraphs. The translation will be clearer in Sec. II.

B. Notations

For a set A and an integer $m \leq |A|$, we define $\binom{A}{m} := \{B \subset A \ : \ |B| = m\}$. Let $[n]$ denote $\{1, \ldots, n\}$ for a positive integer n. Let $\mathbf{e}_i$ be the i^{th} standard unit vector. We use $\mathbb{I}\{\cdot\}$ to denote an indicator function. We shall use $\log(\cdot)$ to indicate the natural logarithm.

II. Problem Formulation

We first present the sampling-based subspace clustering. In this paper, we focus on the case of two subspaces. Consider a collection of n points in an Euclidean space, each of which is either from subspace S_0 or S_1. For an integer $d \geq 2$, a *similarity measure* between d data points is a quantity that tends to be 1 if the d points are from the same subspace, and 0 otherwise. Among all possible d-tuples of data points, similarity measures are accessible for a sampled subcollection. The goal is to cluster the n data points into the two clusters using the sampled similarity measures.

We now formally translate the sampling-based subspace clustering into community recovery in hypergraphs. The i-th point is viewed as node i, and its membership X_i denotes which cluster the data point belongs to, i.e., $X_i = 0$ if the data point is from S_0; $X_i = 1$ otherwise. Let $\mathbf{X} := \{X_i\}_{1 \leq i \leq n}$ be the *ground truth vector*. For a d-tuple $E = \{i_1, i_2, \ldots, i_d\}$, we define a noisy similarity measure Y_E by

$$Y_E = \mathbb{I}\{X_{i_1} = \cdots X_{i_d}\} \oplus Z_E, \tag{1}$$

where $\oplus$ denotes modulo-2 addition, and $Z_E \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ for a fixed noise rate $\theta \in (0, 1/2)$. We assume that similarities measures are collected from a random hypergraph $\mathcal{H} = ([n], \mathcal{E})$ where each element in $\binom{[n]}{d}$ belongs to $\mathcal{E}$ independently with probability $p \in [0, 1]$. We simply write

$\mathcal{H} \sim \mathcal{H}_{n,p}^d$ for this random hypergraph model. *Sample complexity* is defined as the expected number of hyperedges in a random measurement hypergraph, which is $p\binom{n}{d}$. Let $\mathbf{Y} := \{Y_E\}_{E\in\mathcal{E}}$. Then, the sampling-based subspace clustering problem is equivalent to the problem of recovering $\mathbf{X}$ based on $\mathbf{Y}$. We remark that when $d = 2$, this reduces to a conventional graph setting [10]–[13].

Note that given $\mathcal{H}$, the conditional distribution of $\mathbf{Y}|\mathbf{X}$ is equal to that of $\mathbf{Y}|\mathbf{X}\oplus\mathbf{1}$, implying that there is no way to distinguish $\mathbf{X}$ from $\mathbf{X}\oplus\mathbf{1}$. Hence, we allow for a global shift in reconstructing $\mathbf{X}$. Specifically, given a recovery scheme ψ, the probability of error is defined as

$$P_e(\psi) := \max_{\mathbf{X}\in\{0,1\}^n} \Pr\left(\psi(\mathbf{Y}) \notin \{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\}\right).$$

We intend to characterize the minimum sample complexity, above which there exists a recovery algorithm ψ such that $P_e(\psi) \to 0$ as n tends to infinity.

III. Main Results

Our main contribution lies in characterization of the fundamental limit for exact subspace clustering, stated below.

Theorem 1. *Fix $d \geq 2$ and $\epsilon > 0$. Suppose $\mathcal{H} \sim \mathcal{H}_{n,p}^d$. Then,*

$$\begin{cases} \inf_\psi P_e(\psi) \to 0 & \text{if } \binom{n}{d}p \geq (1+\epsilon)\frac{2^{d-2}}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}, \\ \inf_\psi P_e(\psi) \not\to 0 & \text{if } \binom{n}{d}p \leq (1-\epsilon)\frac{2^{d-2}}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}. \end{cases}$$

See Sec. IV for the proof.

While [9] considers a general similarity measurement, it provides only a sufficient condition on sample complexity for weak recovery, which allows a vanishing fraction of misclassified nodes [15]. Specialized to the model of our interest, the sufficient condition in [9] reads $\Omega(n\log^2 n)$, which comes with an extra $\log n$ to the fundamental limit characterized in the above theorem.

One interesting observation in Thm. 1 is that the limit is proportional to $\frac{2^{d-2}}{d}$. One can interpret that the amount of information that one hyperedge can reveal on average is approximately $\frac{d}{2^{d-2}}$. For ease of explanation, suppose that $\theta = 0$ and we observe an hyperedge $E = \{i_1, i_2, \ldots, i_d\}$. The case of $Y_E = 1$ provides $(d-1)$ bits of information as it implies $X_{i_1} = X_{i_2} = \cdots = X_{i_d}$. On the other hand, the case of $Y_E = 0$ provides less information as it rules out only two possible cases ($X_{i_1} = X_{i_2} = \cdots = X_{i_d} = 0$ and $X_{i_1} = X_{i_2} = \cdots = X_{i_d} = 1$) out of 2^d possible candidates. This amounts to roughly $d\cdot\frac{2}{2^d}$ bits. Since $Y_E = 1$ occurs with probability $\frac{1}{2^{d-1}}$, the amount of information that one hyperedge can carry on average should read about $\frac{1}{2^{d-1}}(d-1) + \left(1-\frac{1}{2^{d-1}}\right)\frac{d}{2^{d-1}} \approx \frac{d}{2^{d-2}}$.

Our result also provides a guideline as to how to choose d for sample-efficient subspace clustering. When θ is independent of d, the limit decreases as d decreases. In practical applications, however, θ may change depending on d. The quality of similarity measure across data points can improve with more data points involved, making θ inversely proportional to d. In this case, choosing d as small as possible may not be the best option, i.e., there might be a *sweet spot* for d that minimizes the sample complexity. It turns out this is indeed the case in practice. In Sec. VI, we will discuss in greater details.

IV. Proofs of Information-theoretic Limit

Now we prove the information-theoretic upper and lower bounds; we defer the proofs of lemmas to the full version [16]. We highlight that the probability of error is dependent upon the ground truth vector, and hence a more delicate analysis is needed. We begin with some notations: for a vector $\mathbf{V} = \{V_i\}_{1\leq i\leq n} \in \{0,1\}^n$ and $E = \{i_1, i_2, \ldots, i_d\} \in \binom{[n]}{d}$, let $I_E(\mathbf{V}) = \mathbb{I}\{V_{i_1} = V_{i_2} = \ldots V_{i_d}\}$, and let $\mathbf{I}(\mathbf{V}) = \{I_E(\mathbf{V})\}_{E\in\mathcal{E}}$; define $d(\mathbf{V}) := \|\mathbf{Y} - \mathbf{I}(\mathbf{V})\|_1$.

A. Proof of Achievability

We denote by ψ_{ML} the maximum likelihood (ML) decoder for the problem. One can easily verify that $\psi_{\text{ML}}(\mathbf{Y}) = \arg\min_{\mathbf{V}\in\{0,1\}^n} d(\mathbf{V})$, where ties are randomly broken. We will prove that $\max_{\mathbf{X}\in\{0,1\}^n}\Pr(\psi_{\text{ML}}(\mathbf{Y}) \notin \{\mathbf{X}, \mathbf{X}\oplus\mathbf{1}\}) \to 0$ under the sufficient condition.

Without loss of generality, let $\mathbf{A} \in \{0,1\}^n$ be a ground truth vector whose first k coordinates are 0 and the next $n-k$ coordinates are 1. By symmetry, we assume $0 \leq k \leq n/2$. To list all possible candidates for the ML solution, we denote by $\mathcal{A}_{i,j}$ the collection of vectors which differ from $\mathbf{A}$ in i coordinates among the first k coordinates and in j coordinates among the next $n-k$ coordinates. In particular, $\mathcal{A}_{0,0} = \{\mathbf{A}\}$ and $\mathcal{A}_{k,n-k} = \{\mathbf{A}\oplus\mathbf{1}\}$. Note that

$$\{\mathbf{V}\in\{0,1\}^n\ :\ \mathbf{V}\notin\{\mathbf{A},\mathbf{A}\oplus\mathbf{1}\}\} = \bigcup_{(i,j)\in\mathcal{I}} \mathcal{A}_{i,j},$$

where $\mathcal{I} := \{(i,j)\ :\ (i,j)\notin\{(0,0),(k,n-k)\},\ 0\leq i\leq k,\ \text{and } 0\leq j\leq n-k\}$.

Hence, we get:

$$\begin{aligned} &\Pr(\psi_{\text{ML}}(\mathbf{Y})\notin\{\mathbf{X},\mathbf{X}\oplus\mathbf{1}\} \mid \mathbf{X}=\mathbf{A}) \\ &\overset{(a)}{\leq} \sum_{(i,j)\in\mathcal{I}}\ \sum_{\mathbf{V}\in\mathcal{A}_{i,j}} \Pr\left(d(\mathbf{V})\leq d(\mathbf{A})\right) \\ &= \sum_{(i,j)\in\mathcal{I}} \binom{k}{i}\binom{n-k}{j}\Pr\left(d(\mathbf{V}_{i,j})\leq d(\mathbf{A})\right), \end{aligned} \tag{2}$$

where $\mathbf{V}_{i,j} = (\underbrace{0,\ldots,0}_{i},\underbrace{1,\ldots,1}_{k-i},\underbrace{1,\ldots,1}_{j},\underbrace{0,\ldots,0}_{n-k-j})$ and (a) follows due to the union bound and the fact that ML decoder outputs $\mathbf{V}\notin\{\mathbf{A},\mathbf{A}\oplus\mathbf{1}\}$ only if $d(\mathbf{V})\leq d(\mathbf{A})$.

To compare $d(\mathbf{V}_{i,j})$ and $d(\mathbf{A})$, we define $\mathcal{F}_{i,j} := \{E\in\binom{[n]}{d}\ :\ I_E(\mathbf{A})\neq I_E(\mathbf{V}_{i,j})\}$ and $\mathcal{E}_{i,j} := \mathcal{E}\cap\mathcal{F}_{i,j}$. By definition, for $E\in\mathcal{E}_{i,j}$, $Y_E = I_E(\mathbf{A})$ if $Z_E = 0$; $Y_E = I_E(\mathbf{V}_{i,j})$ otherwise. It follows that $d(\mathbf{V}_{i,j})\leq d(\mathbf{A})$ if and only if $\sum_{E\in\mathcal{E}_{i,j}} Z_E \geq \frac{|\mathcal{E}_{i,j}|}{2}$. Thus,

$$\begin{aligned} &\Pr\left(d(\mathbf{V}_{i,j})\leq d(\mathbf{A})\right) \\ &= \sum_{\ell=1}^{|\mathcal{F}_{i,j}|}\Pr\left(d(\mathbf{V}_{i,j})\leq d(\mathbf{A})\mid|\mathcal{E}_{i,j}|=\ell\right)\Pr(|\mathcal{E}_{i,j}|=\ell) \\ &\overset{(a)}{\leq} \sum_{\ell=1}^{|\mathcal{F}_{i,j}|} e^{-\ell D(0.5\|\theta)}\binom{|\mathcal{F}_{i,j}|}{\ell}p^\ell(1-p)^{|\mathcal{F}_{i,j}|-\ell} \end{aligned} \tag{3}$$

where (a) is due to Chernoff-Hoeffding [17] and $D(p\|q)$ be the Kullback-Leibler (KL) divergence between Bernoulli(p) and Bernoulli(q). (3) can be simplified as

$$(1-(1-e^{-D(0.5\|\theta)}))^{|\mathcal{F}_{i,j}|},$$

hence by letting $p'=(1-e^{-D(0.5\|\theta)})p$, we get

$$(2) \le \sum_{(i,j)\in\mathcal{I}} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|}. \tag{4}$$

To give a tight upper bound on (4), we need to estimate $|\mathcal{F}_{i,j}|$. To this end, we first count $|\mathcal{F}_{i,j}|$. For an exhaustive counting, we define $\mathcal{B}_1 := \{1,2,\ldots,i\}$, $\mathcal{B}_2 := \{i+1, i+2,\ldots,k\}$, $\mathcal{B}_3 := \{k+1,k+2,\ldots,k+j\}$, and $\mathcal{B}_4 := \{k+j+1,k+j+2,\ldots,n\}$. Note that

- $I_E(\mathbf{A})=1$ and $I_E(\mathbf{V}_{i,j})=0$ when either
  (Case 1) $E\subset(\mathcal{B}_1\cup\mathcal{B}_2),\ E\cap\mathcal{B}_1\neq\emptyset,\ E\cap\mathcal{B}_2\neq\emptyset$, or
  (Case 2) $E\subset(\mathcal{B}_3\cup\mathcal{B}_4),\ E\cap\mathcal{B}_3\neq\emptyset,\ E\cap\mathcal{B}_4\neq\emptyset$;
- $I_E(\mathbf{A})=0$ and $I_E(\mathbf{V}_{i,j})=1$ when either
  (Case 3) $E\subset(\mathcal{B}_1\cup\mathcal{B}_4),\ E\cap\mathcal{B}_1\neq\emptyset,\ E\cap\mathcal{B}_4\neq\emptyset$, or
  (Case 4) $E\subset(\mathcal{B}_2\cup\mathcal{B}_3),\ E\cap\mathcal{B}_2\neq\emptyset,\ E\cap\mathcal{B}_3\neq\emptyset$.

This yields:

$$|\mathcal{F}_{i,j}| = \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{k-i}{d-\ell}+\sum_{\ell=1}^{d-1}\binom{j}{\ell}\binom{n-k-j}{d-\ell} + \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n-k-j}{d-\ell}+\sum_{\ell=1}^{d-1}\binom{k-i}{\ell}\binom{j}{d-\ell}. \tag{5}$$

Observe that $|\mathcal{F}_{i,j}|=|\mathcal{F}_{k-i,n-k-j}|$, which implies (i,j)-th term in the summation (4) is equal to $(k-i,n-k-j)$-th term. Thus, one can find an one-to-one mapping between $\{(i,j)\in\mathcal{I}\ :\ j\le\lfloor\frac{n-k}{2}\rfloor\}$ and $\{(i,j)\in\mathcal{I}\ :\ j\ge\lceil\frac{n-k}{2}\rceil\}$ so that corresponding terms have the same value. By using this mapping, we can manipulate:

$$(4) \quad \le 2\sum_{(i,j)\in\mathcal{I},j\le\frac{n-k}{2}}\binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|}. \tag{6}$$

In this regard, we only consider the case of $j\le\frac{n-k}{2}$. Let $\delta>0$ be a small constant. For the case where $i\ge\delta n$ or $j\ge\delta n$, it follows from (5) that

$$|\mathcal{F}_{i,j}| \ge \sum_{\ell=1}^{d-1}\binom{j}{\ell}\binom{n-k-j}{d-\ell}+\sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n-k-j}{d-\ell} \overset{(a)}{\ge} \sum_{\ell=1}^{d-1}\binom{j}{\ell}\binom{n/4}{d-\ell}+\sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n/4}{d-\ell} \overset{(b)}{=} \Omega(n^d),$$

where (a) follows since $j\le\frac{n-k}{2}$ and $k\le\frac{n}{2}$; (b) follows since $i\ge\delta n$ or $j\ge\delta n$. Then it is easy to show (6)$\to 0$ for this case:

$$(6)\cdot\mathbb{I}\{i\ge\delta n \text{ or } j\ge\delta n\} \le 2\sum_{(i,j)\in\mathcal{I}}\binom{k}{i}\binom{n-k}{j}e^{-p'\Omega(n^d)}$$
$$= 2e^{-\Omega(n\log n)}\sum_{(i,j)\in\mathcal{I}}\binom{k}{i}\binom{n-k}{j} \le 2e^{-\Omega(n\log n)}2^n\to 0.$$

Now we are left with the case where $i<\delta n$ and $j<\delta n$. The following lemma gives a tight lower bound on $|\mathcal{F}_{i,j}|$ in this case:

Lemma 1. *For $i<\delta n$ and $j<\delta n$,*

$$|\mathcal{F}_{i,j}| \ge (i+j)\cdot\frac{(1-2\delta)^{d-1}}{2^{d-2}}\binom{n}{d-1}.$$

By Lem. 1,

$$\begin{aligned}
&\frac{1}{2}\cdot(6)\cdot\mathbb{I}\{i<\delta n \text{ and } j<\delta n\}\\
&\le \sum_{i<\delta n,j<\delta n}\mathbb{I}\{i\neq 0 \text{ or } j\neq 0\}\binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|}\\
&\le \sum_{i<\delta n,j<\delta n}\mathbb{I}\{i\neq 0 \text{ or } j\neq 0\}n^in^je^{-p'(i+j)\cdot\frac{(1-2\delta)^{d-1}}{2^{d-2}}\binom{n}{d-1}}\\
&\le \sum_{i<\delta n,j<\delta n}\mathbb{I}\{i\neq 0 \text{ or } j\neq 0\}\\
&\qquad\cdot\exp\left((i+j)\left\{\log n-\frac{p'(1-2\delta)^{d-1}\binom{n}{d-1}}{2^{d-2}}\right\}\right).
\end{aligned}\tag{7}$$

As $(1-e^{-D(0.5\|\theta)})=(\sqrt{1-\theta}-\sqrt{\theta})^2$, the precondition in Thm. 1 becomes

$$\binom{n}{d}p' \ge (1+\epsilon)\frac{2^{d-2}}{d}n\log n, \tag{8}$$

and hence we get:

$$\begin{aligned}
\frac{p'(1-2\delta)^{d-1}\binom{n}{d-1}}{2^{d-2}} &\ge (1+\epsilon)(1-2\delta)^{d-1}\log n\\
&\ge (1+\epsilon/2)\log n,
\end{aligned}$$

where the last inequality follows by choosing δ sufficiently small. Thus, (7) converge to zero as n tends to infinity.

B. Proof of Converse

Without loss of generality, we assume that n is even. Let $\mathcal{V}_{1/2}$ be the collection of n-dimensional vectors which consist of $n/2$ coordinates of 0's and $n/2$ coordinates of 1's. Moreover, let $\mathbf{X}_{1/2}$ be the random vector taking a sample uniformly over $\mathcal{V}_{1/2}$. For any scheme ψ, by definition of $P_e(\psi)$, we see that

$$\Pr\left(\psi(\mathbf{Y})\notin\{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\}\mid\mathbf{X}=\mathbf{X}_{1/2}\right)\le P_e(\psi)$$

and hence

$$\inf_\psi\Pr\left(\psi(\mathbf{Y})\notin\{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\}\mid\mathbf{X}=\mathbf{X}_{1/2}\right)\le\inf_\psi P_e(\psi).$$

On the other hand, one can easily deduce that the infimum at the left hand side is achieved by $\psi_{\mathrm{ML},1/2}$ which is defined as $\psi_{\mathrm{ML},1/2}(\mathbf{Y})=\arg\min_{\mathbf{V}\in\mathcal{V}_{1/2}}d(\mathbf{V})$, where ties are broken randomly. Let $\mathbf{A}=(\underbrace{0,\ldots,0}_{n/2},\underbrace{1,\ldots,1}_{n/2})$. By symmetry,

$$\begin{aligned}
&\Pr\left(\psi_{\mathrm{ML},1/2}(\mathbf{Y})\notin\{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\}\mid\mathbf{X}=\mathbf{X}_{1/2}\right)\\
&=\Pr\left(\psi_{\mathrm{ML},1/2}(\mathbf{Y})\notin\{\mathbf{A},\ \mathbf{A}\oplus\mathbf{1}\}\mid\mathbf{X}=\mathbf{A}\right).
\end{aligned}$$

Let S be the success event that $\arg\min_{\mathbf{V}\in\mathcal{V}_{1/2}} d(\mathbf{V}) = \{\mathbf{A}, \mathbf{A}\oplus\mathbf{1}\}$ when $\mathbf{X}=\mathbf{A}$. Notice that

$$\Pr\left(\psi_{\mathrm{ML},1/2}(\mathbf{Y}) \notin \{\mathbf{A},\ \mathbf{A}\oplus\mathbf{1}\} \mid \mathbf{X}=\mathbf{A}\right) \overset{(a)}{\geq} \frac{1}{3}\Pr(S^c),$$

where (a) follows since given S^c, there are more than two candidates for $\arg\min_{\mathbf{V}\in\mathcal{V}_{1/2}} d(\mathbf{V})$, so

$$\Pr\left(\psi_{\mathrm{ML},1/2}(\mathbf{Y}) \notin \{\mathbf{A},\ \mathbf{A}\oplus\mathbf{1}\} \mid \mathbf{X}=\mathbf{A},\ S^c\right) \geq \frac{1}{3}.$$

Hence, it suffices to show $\Pr(S)\to 0$. By definition,

$$\Pr(S) = \Pr\left(\bigcap_{\mathbf{V}\in\mathcal{V}_{1/2}\setminus\{\mathbf{A},\mathbf{A}\oplus\mathbf{1}\}} [d(\mathbf{V}) > d(\mathbf{A})]\right). \tag{9}$$

To give an upper bound tight enough, we construct a large subset of nodes such that any two nodes in the subset do not share the same hyperedge. To this end, we first choose a big subset $\mathcal{R}_{\text{big}} = \{1,2,\dots,r\}\cup\{\frac{n}{2}+1,\frac{n}{2}+2,\dots,\frac{n}{2}+r\}$, where $r=\lceil\frac{n}{\log^3 n}\rceil$. Then erase every node in $\mathcal{R}_{\text{big}}$ which shares hyperedges with other nodes in $\mathcal{R}_{\text{big}}$ to obtain $\mathcal{R}_{\text{res}}$. The following lemma guarantees that $\mathcal{R}_{\text{res}}$ has a comparable size with $\mathcal{R}_{\text{big}}$ with high probability:

Lemma 2. *Suppose* $\binom{n}{d}p = O(n\log n)$. *Then with probability approaching* 1,

$$|\mathcal{R}_{res}| \geq (2-o(1))r.$$

Let Δ be the event that $|\mathcal{R}_{\text{res}}| \geq (2-o(1))r$. Given the event Δ,

$$\{1,2,\cdots,n/2\}\cap\mathcal{R}_{\text{res}} \tag{10}$$

and

$$\left\{\frac{n}{2}+1,\frac{n}{2}+2,\cdots,n\right\}\cap\mathcal{R}_{\text{res}} \tag{11}$$

should contain more than $r/2$ elements; we collect $r/2$ elements from (10) (resp. (11)) and denote by $\{b_1,b_2,\dots,b_{r/2}\}$ (resp. $\{c_1,c_2,\dots,c_{r/2}\}$).

Now we give an upper bound on (9). Suppose that there exist k,ℓ such that $d(\mathbf{A}\oplus\mathbf{e}_{b_k})\leq d(\mathbf{A})$ and $d(\mathbf{A}\oplus\mathbf{e}_{c_\ell})\leq d(\mathbf{A})$. Conditioning on Δ, there are no hyperedges that contain both b_k and c_ℓ, so $d(\mathbf{A}\oplus\mathbf{e}_{b_k}\oplus\mathbf{e}_{c_\ell})\leq d(\mathbf{A})$. From this observation, we see that conditioning on Δ,

$$\begin{aligned} S &\subset \bigcap_{k=1}^{r/2}[d(\mathbf{A}\oplus\mathbf{e}_{b_k})>d(\mathbf{A})]\cup\bigcap_{k=1}^{r/2}[d(\mathbf{A}\oplus\mathbf{e}_{c_k})>d(\mathbf{A})] \\ &=: S'. \end{aligned}$$

Hence, we get:

$$\begin{aligned} \Pr(S) &\overset{(a)}{\lesssim} \Pr\left(S' \mid \Delta\right) \\ &\leq 2\Pr\left(\bigcap_{k=1}^{r/2}[d(\mathbf{A}\oplus\mathbf{e}_{b_k})>d(\mathbf{A})] \;\middle|\; \Delta\right) \\ &\overset{(b)}{=} 2\Pr\left(d(\mathbf{A}\oplus\mathbf{e}_{b_1})>d(\mathbf{A}) \mid \Delta\right)^{r/2}, \end{aligned}$$

where (a) is due to Lem. 2; (b) follows from the fact that events $\{[d(\mathbf{A}\oplus\mathbf{e}_{b_k})>d(\mathbf{A})\mid\Delta]\}_{1\leq k\leq r/2}$ are mutually independent. Now we finish the proof. Let $p'=(1-e^{-D(0.5\|\theta)})p$ as in the achievability proof.

Lemma 3.

$$\Pr\left(d(\mathbf{A}\oplus\mathbf{e}_{b_1})\leq d(\mathbf{A}) \;\middle|\; \Delta\right) \geq (1+o(1))e^{-2p'\binom{n/2-1}{d-1}}.$$

By Lem. 3, we get:

$$\begin{aligned} &\Pr\left(d(\mathbf{A}\oplus\mathbf{e}_{b_1})>d(\mathbf{A}) \mid \Delta\right)^{r/2} \\ &\leq \left(1-(1+o(1))e^{-2p'\binom{n/2-1}{d-1}}\right)^{r/2} \\ &\leq \exp\left(-(1+o(1))\frac{r}{2}\exp\left\{-2p'\binom{n/2-1}{d-1}\right\}\right) \\ &\leq \exp\left(-(1+o(1))\frac{n}{2\log^3 n}\exp\left\{-(1+o(1))\cdot\frac{p'd\binom{n}{d}}{2^{d-2}n}\right\}\right), \end{aligned}$$

and the last term converges to zero as $p'\leq(1-\epsilon)\frac{2^{d-2}}{d}\frac{n\log n}{\binom{n}{d}}$.

V. An Efficient Algorithm

In this section, we present an algorithm that achieves the information-theoretic limit characterized in Thm. 1. We only describe the algorithm while deferring a detailed analysis to the full version [16]. The algorithm operates in two stages, beginning with a decent initial estimate from spectral clustering followed by iterative refinement. Detailed procedures are presented in Alg. 1. Our algorithm is inspired by similar two-stage approaches that have been applied to many other problems [10], [13], [18].

Algorithm 1

1: For given $\mathbf{Y}=\{Y_E\}_{E\in\mathcal{E}}$, construct an $n\times n$ matrix

$$\mathbf{A} := \sum_{E\in\mathcal{E}}\sum_{\substack{\{i,j\}\subset E\\ i\neq j}} Y_E\left\{\mathbf{e}_i\mathbf{e}_j^T+\mathbf{e}_j\mathbf{e}_i^T\right\}.$$

2: Find the first two eigenvectors of $\mathbf{A}$ and stack them side by side to obtain $\mathbf{U}\in\mathbb{R}^{n\times 2}$.

3: Cluster the rows of $\mathbf{U}$ using the approximate geometric k-clustering [19] with any fixed approximation rate $\epsilon>0$.

4: Classify nodes according to the resulting clustering to obtain the membership vector $\mathbf{X}^{(0)}=\{X_i^{(0)}\}_{1\leq i\leq n}\in\{0,1\}^n$.

5: For $T=O(\log n)$, update $\mathbf{X}^{(t)}=\{X_i^{(t)}\}_{1\leq i\leq n}$ for $t=0,1,\dots,T-1$ by

$$X_i^{(t+1)} = \begin{cases} X_i^{(t)} & \text{if } d(\mathbf{X}^{(t)})<d(\mathbf{X}^{(t)}\oplus\mathbf{e}_i), \\ X_i^{(t)}\oplus 1 & \text{if } d(\mathbf{X}^{(t)})\geq d(\mathbf{X}^{(t)}\oplus\mathbf{e}_i), \end{cases}$$

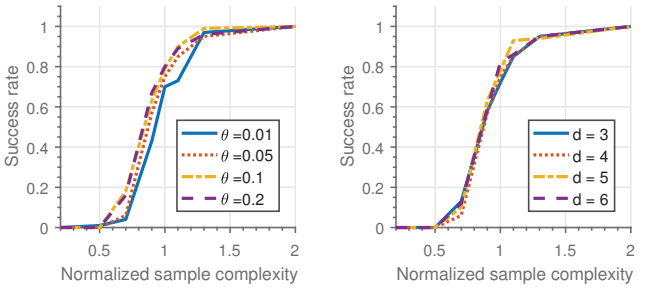
for $i=1,2,\dots,n$, where $d(\cdot)$ is defined in Sec. IV.

6: Output $\mathbf{X}^{(T)}=\{X_i^{(T)}\}_{1\leq i\leq n}$.

VI. Discussion

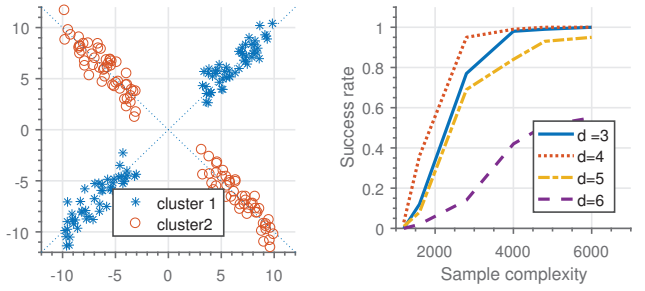
A. Information-theoretic Limits

In this section, we demonstrate the performance of Alg. 1 by running Monte Carlo simulations. Each point plotted in



(a) Varying θ (b) Varying d

Fig. 1: **Alg. 1 achieves the optimal sample complexity.** We run Monte Carlo simulations to estimate the probability of success when: (a) $n = 1000$, $d = 4$, and for various choices of θ; (b) $n = 1000$, $\theta = 0.05$, and for various choices of d. For each curve, we normalize the number of samples by the respective information theoretic limits, characterized in Thm. 1. Observe that the probability of success quickly approaches 1 as the normalized sample complexity crosses 1.



(a) Randomly generated data set (b) Varying d

Fig. 2: **Optimal choice of d when θ decays with d.** We run Monte Carlo simulations to estimate the probability of success with the data set shown in (a). We observe that the effective noise rate decreases as d increases. For varying d from 3 to 6, the success probability of Alg. 1 is shown in (b): the best performance of the algorithm is observed when $d = 4$.

Fig. 1a and Fig. 1b indicates an empirical success rate. We take 100 Monte Carlo trials. Fig. 1a shows the probability of success when $n = 1000$, $d = 4$, and for various choices of θ. Shown in Fig. 1b is the performance of our algorithm with $n = 1000$, $\theta = 0.05$, and for various choices of d. For both figures, the x-axis denotes the number of samples normalized by the respective information-theoretic limits, characterized in Thm. 1. One can observe that the success probability due to Alg. 1 quickly approaches 1 as the normalized sample complexity crosses 1, which corroborates our theoretical findings.

B. Optimal d for Subspace Clustering

In this section, we observe how the fundamental limit varies as a function of d. As we briefly discussed in Sec. III, if the noise rate θ is irrelevant to d, the optimal choice of d would be the minimum possible value of d. However, if the noise quality θ depends on d, there may be a sweet spot for d. In practice, the noise quality can be highly dependent on d. To see this, consider a synthetic data set for subspace clustering, shown in Fig. 2a. Here the goal is to cluster n $(= 200)$ 2-dimensional data points approximately lying on a union of two lines (1-dimensional subspaces). For each sampled hyperedge $E = \{i_1, \cdots, i_d\}$, we evaluate Y_E in (1) as follows: we first find the line that minimizes the sum of distances to the d points $i_1, \cdots i_d$, and compute the sum of distances from the d points to the line that we denote by D. Then, we set $Y_E = 1$ if and only if D is less than a fixed threshold, which is appropriately chosen so that $\Pr(Y_E = 0 \mid i_1, i_2, \cdots, i_d$ are from the same line$) \approx \Pr(Y_E = 1 \mid i_1, i_2, \cdots, i_d$ are not from the same line$)$. We estimate the effective noise rate $\hat{\theta} := \Pr(Y_E = 0 \mid i_1, i_2, \cdots, i_d$ are from the same line$)$ for various d, and observe that $\hat{\theta}$ quickly decreases as d increases.

Using these similarity measures and the synthetic data set, we evaluate the performance of Alg. 1, shown in Fig. 2b. Interestingly, we find that the optimal choice of d here is 4 rather than 3.

REFERENCES

[1] R. Vidal, R. Tron, and R. Hartley, "Multiframe motion segmentation with missing data using powerfactorization and gpca," *International Journal of Computer Vision*, vol. 79, no. 1, pp. 85–105, 2008.

[2] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on*, vol. 1, 2003.

[3] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, March 2011.

[4] G. Chen and G. Lerman, "Spectral curvature clustering (scc)," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 317–330, 2009.

[5] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[6] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk, "Greedy feature selection for subspace clustering." *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2487–2517, 2013.

[7] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6320–6342, 2015.

[8] V. M. Govindu, "A tensor decomposition for geometric grouping and segmentation," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005.

[9] D. Ghoshdastidar and A. Dukkipati, "Uniform hypergraph partitioning: Provable tensor methods and sampling techniques," *CoRR*, vol. abs/1602.06516, 2016.

[10] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2016.

[11] E. Mossel, J. Neeman, and A. Sly, "Stochastic block models and reconstruction," *arXiv preprint arXiv:1202.1499*, 2012.

[12] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2788–2797, 2016.

[13] Y. Chen, G. Kamath, C. Suh, and D. Tse, "Community recovery in graphs with locality," in *International Conference on Machine Learning*, 2016.

[14] K. Ahn, K. Lee, and C. Suh, "Community recovery in hypergraphs," in *Allerton Conference on Communication, Control and Computing*, 2016.

[15] B. Hajek, Y. Wu, and J. Xu, "Information limits for recovering a hidden community," in *Information Theory (ISIT), 2016 IEEE International Symposium on*, 2016.

[16] K. Ahn, K. Lee, and C. Suh, "Information-theoretic limits of subspace clustering (full version)," https://sites.google.com/site/kw1jjang/AhnLeeSuh2017full.pdf, 2017.

[17] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[18] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Advances in Neural Information Processing Systems*, 2015.

[19] J. Matoušek, "On approximate geometric k-clustering," *Discrete & Computational Geometry*, vol. 24, no. 1, pp. 61–84, 2000.