On the uniqueness and stability of dictionaries for sparse representation of noisy signals

Charles J. Garfinkle and Christopher J. Hillar
Redwood Center for Theoretical Neuroscience, Berkeley, CA, USA

***Abstract*—Learning optimal dictionaries for sparse coding has exposed characteristic sparse features of many natural signals. However, universal conditions guaranteeing the consistency of these learned features in the presence of measurement noise have yet to be stated. Here, we prove very generally that optimal dictionaries and sparse codes are uniquely and stably determined up to measurement error for all diverse enough datasets generated by the sparse coding model. Applications are given to data analysis, engineering, and neuroscience.**

I. Introduction

A common assumption underlying contemporary solutions to problems in signal processing and pattern analysis is that each of N observed n-dimensional signal samples is a (noisy) linear combination of at most k elementary waveforms drawn from some unknown "dictionary" of size m, typically with $m \ll N$ (see [**?**] for a comprehensive review of this and related models). Approximate solutions to this sparsity-constrained inverse problem and related problems have revealed seemingly characteristic structure in signal classes lacking idiosyncratic formal models (e.g., in vision [**?**]). Interestingly, a seminal work in the field demonstrated that dictionaries optimized with respect to small patches of natural images share qualitative similarities with filters fit to the response properties of simple-cell neurons in mammalian visual cortex [**?**], [**?**], [**?**], [**?**], a major result in theoretical neuroscience. Moreover, these latent waveforms (e.g., "Gabor" wavelets) estimated from data appear to be canonical [**?**]; i.e. they appear in dictionaries learned by a variety of algorithms trained with respect to different natural image datasets.

Motivated by these discoveries and earlier work which proposed a theory of communication between sparsely active neural populations contingent on uniqueness [**?**], we address when dictionaries and the sparse representations they induce are uniquely determined by data. Answers to this question have other practical implications. For example, it has been demonstrated that a sparse coding analysis of painting style can detect forgeries [**?**]; but this implies all dictionaries consistent with training data generalize in their ability to sparsely encode new samples. Fortunately, several algorithms have recently been proposed which provably recover latent dictionaries under idiosyncratic conditions (see [**?**, Sec. I-E] for a summary of the state-of-the-art). Few theorems, however, can be cited to explain the consistency of inference under this model of data more broadly; in particular, a universal guarantee of uniqueness and stability in the context of noise has yet to emerge.

Here, we prove very generally that uniqueness and stability of solutions is an expected property of the sparse coding problem described above as well as a related (and perhaps more well-known) optimization problem which asks for a dictionary and set of sparse codes that minimize the mean sparsity of the codes over all data samples.

Formally, let $\mathbf{A}$ be a real $n \times m$ matrix with columns $\mathbf{A}_j$ $(j = 1, \ldots, m)$ and let dataset Z consist of measurements:

$$\mathbf{z}_i = \mathbf{A}\mathbf{x}_i + \mathbf{n}_i, \quad i = 1, \ldots, N, \tag{1}$$

for *k-sparse* $\mathbf{x}_i \in \mathbb{R}^m$ having at most $k < m$ nonzero entries and *noise* $\mathbf{n}_i \in \mathbb{R}^n$, with bounded norm $\|\mathbf{n}_i\|_2 \leq \eta$ representing our combined worst-case uncertainty in measuring $\mathbf{A}\mathbf{x}_i$. We show that dictionaries $\mathbf{A}$ injective on certain subsets of k-sparse codes are identifiable from as few as $N = m(k-1)\binom{m}{k} + m$ noisy sparse linear combinations of their columns up to an error that is linear in the noise (Thm. 1). Specifically, the first mathematical problem we consider is:

Problem 1. *Find a real $n \times \overline{m}$ matrix $\mathbf{B}$ and k-sparse $\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_N$ satisfying $\|\mathbf{z}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \eta$ for all i.*

Our claim is that there exist very general conditions under which any such matrix $\mathbf{B}$ always contains a submatrix "close" to some submatrix of $\mathbf{A}$, and similarly for the sparse codes $\overline{\mathbf{x}}_i$ and $\mathbf{x}_i$. In fact, provided $n \geq \min(2k, m)$, in almost all cases 1 is well-posed (as per Hadamard [**?**]) given enough data (Cor. 2). Importantly, these explicit, universal guarantees apply without the imposition of *any* constraints on the recovered matrix $\mathbf{B}$ (e.g. that it, too, satisfy any injectivity criteria) beyond an upper bound on its number of columns $\overline{m}$.

We note that any particular solution to Problem 1 actually represents an infinity of equivalent solutions $\mathbf{BPD}$ and $\mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_1, \ldots, \mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_N$ parametrized by a choice of permutation matrix $\mathbf{P}$ and invertible diagonal matrix $\mathbf{D}$ capturing the sparsity-preserving labeling and scaling ambiguities inherent to the problem statement. Previous theoretical work addressing the noiseless case $\eta = 0$ (e.g., [**?**], [**?**], [**?**], [**?**]) with fixed $\overline{m} = m$ has shown that the solution to Prob. 1, when it exists, is unique up to these ambiguities provided the $\mathbf{x}_i$ are sufficiently diverse and the matrix $\mathbf{A}$ satisfies:

$$\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_2 \implies \mathbf{x}_1 = \mathbf{x}_2, \quad \text{for all } k\text{-sparse } \mathbf{x}_1, \mathbf{x}_2, \tag{2}$$

also called the *spark condition*, which in any case is necessary to guarantee the uniqueness of arbitrary sequences of k-sparse $\mathbf{x}_i$. We generalize these results to the practical setting of additive noise by incorporating a notion of stability with respect to noise.

Definition 1. *Fix $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_N\} \subset \mathbb{R}^n$. We say Y has a **k-sparse representation in** $\mathbb{R}^m$ if there exists a matrix $\mathbf{A}$ and k-sparse $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^m$ such that $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$ for all i. This representation is **stable** if for every $\delta_1, \delta_2 \geq 0$, there exists some $\varepsilon = \varepsilon(\delta_1, \delta_2)$ that is strictly positive for positive δ_1 and δ_2 such that if $\mathbf{B}$ and k-sparse $\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_N \in \mathbb{R}^m$ satisfy:*

$$\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \varepsilon(\delta_1, \delta_2), \quad \text{for all } i = 1, \dots, N,$$

then there is some permutation matrix $\mathbf{P}$ and invertible diagonal matrix $\mathbf{D}$ such that for all i, j:

$$\|\mathbf{A}_j - \mathbf{B}\mathbf{P}\mathbf{D}_j\|_2 \leq \delta_1 \quad \text{and} \quad \|\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top \overline{\mathbf{x}}_i\|_1 \leq \delta_2. \quad (3)$$

To see how Def. 1 is motivated by Prob. 1, suppose that Y has a stable k-sparse representation in $\mathbb{R}^m$ and fix δ_1, δ_2 to be the desired recovery accuracy in (3). Now, consider any dataset Z generated as in (1) with $\eta \leq \frac{1}{2}\varepsilon(\delta_1, \delta_2)$. From the triangle inequality, it follows that any $n \times m$ matrix $\mathbf{B}$ and k-sparse $\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_N$ solving Prob. 1 are necessarily within δ_1, δ_2 of the original dictionary $\mathbf{A}$ and codes $\mathbf{x}_i$, respectively.

In the next section, we give precise statements of our main results, which include an explicit form for $\varepsilon(\delta_1, \delta_2)$. We then prove our main theorem (Thm. 1) in Sec. III after stating some additional definitions and lemmas required for the proof, including a useful result in combinatorial matrix analysis (Lem. 1, proven in Appendix). We also provide an argument extending our guarantees to the following more common formulation of the dictionary learning problem as an optimization over model parameters minimizing the total (or equivalently, the mean) number of non-zero entries in the sparse codes (Thm. 2):

Problem 2. *Find real $\mathbf{B}$ and $\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_N$ that solve:*

$$\min \sum_{i=1}^{N} \|\overline{\mathbf{x}}_i\|_0 \quad \text{subject to} \quad \|\mathbf{z}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \eta, \text{ for all } i. \quad (4)$$

We then sketch proofs of extensions of Thm. 1 to data and dictionaries drawn from a probability distribution (Thm. 3 and Cor. 2). Finally, in Sec. IV, we discuss both theoretical and practical applications of our main mathematical findings.

II. Results

Before stating our results precisely, we identify criteria on the support sets of the generating codes $\mathbf{x}_i$ that imply stable sparse representations. Letting $\{1, \dots, m\}$ be denoted $[m]$, its power set $2^{[m]}$, and $\binom{[m]}{k}$ the set of subsets of $[m]$ of size k, we say a hypergraph $\mathcal{H} \subseteq 2^{[m]}$ on vertices $[m]$ is *k-uniform* when $\mathcal{H} \subseteq \binom{[m]}{k}$. The *degree* $\deg_{\mathcal{H}}(i)$ of a node $i \in [m]$ is the number of sets in $\mathcal{H}$ that contain i, and we say $\mathcal{H}$ is *regular* when for some r we have $\deg_{\mathcal{H}}(i) = r$ for all i (given such an r, we say $\mathcal{H}$ is *r-regular*). We also write $2\mathcal{H} := \{S \cup S' : S, S' \in \mathcal{H}\}$.

Definition 2. *Given $\mathcal{H} \subseteq 2^{[m]}$, the **star** $\sigma(i)$ is the collection of sets in $\mathcal{H}$ containing i. We say $\mathcal{H}$ has the **singleton intersection property (SIP)** when $\cap\sigma(i) = \{i\}$ for all $i \in [m]$.*

Next, we describe a quantitative generalization of the spark condition. The *lower bound* of a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ is the largest α with $\|\mathbf{M}\mathbf{x}\|_2 \geq \alpha\|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{R}^m$ [**?**]. By compactness of the unit sphere, injective linear maps have nonzero lower bound; hence, if $\mathbf{M}$ satisfies (2), then each submatrix formed from $2k$ of its columns or less has a strictly positive lower bound.

We generalize this lower bound to a domain-restricted *union of subspaces* model [**?**] derived from a hypergraph $\mathcal{H} \subseteq \binom{[m]}{k}$. Let $\mathbf{M}_S$ denote the submatrix formed by the columns of $\mathbf{M}$ indexed by $S \subseteq [m]$, with $\mathbf{M}_\emptyset := \mathbf{0}$. (In the sections that follow, we write $\boldsymbol{\mathcal{M}}_S$ to denote the column-span of a submatrix $\mathbf{M}_S$, and $\boldsymbol{\mathcal{M}}_{\mathcal{G}}$ to denote $\{\boldsymbol{\mathcal{M}}_S\}_{S \in \mathcal{G}}$.) Define:

$$L_{\mathcal{H}}(\mathbf{M}) := \frac{1}{\sqrt{k}} \min \left\{ \frac{\|\mathbf{M}_S \mathbf{x}\|_2}{\|\mathbf{x}\|_2} : S \in \mathcal{H}, \quad \mathbf{x} \in \mathbb{R}^{|S|} \right\}, \quad (5)$$

where we write L_k when $\mathcal{H} = \binom{[m]}{k}$.[1] Clearly, $L_{2k}(\mathbf{M}) > 0$ for $\mathbf{M}$ satisfying (2) and $L_{k'}(\mathbf{M}) \geq L_k(\mathbf{M})$ whenever $k' \leq k$. Note also that $L_2 \geq L_{2\mathcal{H}} \geq L_{2k}$ if $\mathcal{H} \subseteq \binom{[m]}{k}$ has $\cup\mathcal{H} = [m]$.

A vector $\mathbf{x}$ is said to be *supported* in $S \subseteq [m]$ when $\mathbf{x} \in \text{span}\{\mathbf{e}_j : j \in S\}$, where $\mathbf{e}_j$ form the standard column basis in $\mathbb{R}^m$. A set of k-sparse vectors is said to be in *general linear position* when any k of them are linearly independent.

We are now in a position to give a precise statement of our main result, though we will leave the quantity $C_1 = C_1(\mathbf{A}, \{\mathbf{x}_i\}_{i=1}^N, \mathcal{H})$ undefined until Eq. (15). All of our theorems assume matrices consist of real numbers.

Theorem 1. *Fix an $n \times m$ matrix $\mathbf{A}$ with $L_{2\mathcal{H}}(\mathbf{A}) > 0$ for an r-regular $\mathcal{H} \subseteq \binom{[m]}{k}$ with the SIP. If $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbb{R}^m$ contains, for each $S \in \mathcal{H}$, more than $(k-1)\binom{m}{k}$ k-sparse vectors in general linear position supported in S, then there is $C_1 > 0$ with the following holding for all[2] $\varepsilon < L_2(\mathbf{A})/C_1$:*

Every $n \times \overline{m}$ matrix $\mathbf{B}$ for which there are k-sparse $\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_N$ satisfying $\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \varepsilon$ for all i has $\overline{m} \geq m$ and, provided $(r-1)\overline{m} < mr$,

$$\|\mathbf{A}_j - \mathbf{B}_{\overline{J}}\mathbf{P}\mathbf{D}_j\|_2 \leq C_1\varepsilon, \quad \text{for all } j \in J, \quad (6)$$

for some nonempty $\overline{J} \subseteq [\overline{m}]$ and $J \subseteq [m]$ of size $\overline{m} - r(\overline{m} - m)$, permutation matrix $\mathbf{P}$, and invertible diagonal matrix $\mathbf{D}$.

Moreover, if A satisfies (2) and $\varepsilon < L_{2k}(\mathbf{A})/C_1$, then $L_{2k}(\mathbf{B}\mathbf{P}\mathbf{D}) \geq L_{2k}(\mathbf{A}) - C_1\varepsilon$ and:

$$\|\mathbf{x}_i - \left(\mathbf{D}^{-1}\mathbf{P}^\top\right)_{\overline{J}} \overline{\mathbf{x}}_i\|_1 \leq \left(\frac{1 + C_1\|\mathbf{x}_i\|_1}{L_{2k}(\mathbf{A}) - C_1\varepsilon}\right)\varepsilon, \quad \text{for } i \in [N], \quad (7)$$

where $\mathbf{x}_i$ and $\overline{\mathbf{x}}_i$ here represent subvectors formed from restricting to entries indexed by J and $\overline{J}$, respectively.

In words, Thm. 1 says that the smaller the difference $\overline{m} - m$ between the assumed and actual number of latent dictionary elements, the more columns and coefficients of the original dictionary $\mathbf{A}$ and codes $\mathbf{x}_i$ are contained (up to noise) in the appropriately scaled recovered dictionary $\mathbf{B}$ and codes $\overline{\mathbf{x}}_i$, respectively. In the particular case when $\overline{m} = m$, the theorem

[1] We note that $1 - \sqrt{k}L_k(\mathbf{M})$ is known as the asymmetric lower restricted isometry constant for matrices $\mathbf{M}$ with unit ℓ_2-norm columns [**?**].

[2] The condition $\varepsilon < L_2(\mathbf{A})/C_1$ is necessary; otherwise, with $\mathbf{A} = \mathbf{I}$ and $\mathbf{x}_i = \mathbf{e}_i$, the matrix $\mathbf{B} = \left[\mathbf{0}, \frac{1}{2}(\mathbf{e}_1 + \mathbf{e}_2), \mathbf{e}_3, \dots, \mathbf{e}_m\right]$ and 1-sparse codes $\overline{\mathbf{x}}_i = \mathbf{e}_2$ for $i = 1, 2$ and $\overline{\mathbf{x}}_i = \mathbf{e}_i$ for $i \geq 3$ satisfy $\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \varepsilon$ but nonetheless violate (6).

directly implies that $Y = \{\mathbf{A}\mathbf{x}_1, \dots, \mathbf{A}\mathbf{x}_N\}$ has a stable k-sparse representation in $\mathbb{R}^m$, with inequalities (3) guaranteed for ε in Def. 1 given by:

$$\varepsilon(\delta_1, \delta_2) := \min\left\{ \frac{\delta_1}{C_1}, \frac{\delta_2 L_{2k}(\mathbf{A})}{1 + C_1\left(\delta_2 + \max_{i\in[N]} \|\mathbf{x}_i\|_1\right)} \right\}. \tag{8}$$

Note that sparse codes $\mathbf{x}_i$ with a shared support that are in general linear position are straightforward to produce using a "Vandermonde" matrix construction (i.e., use the columns of the matrix $[\gamma_i^j]_{i,j=1}^{k,N}$, for distinct nonzero γ_i). Thus, the assumptions of Thm. 1 are easily met.

Corollary 1. *Given a regular hypergraph $\mathcal{H} \subseteq \binom{[m]}{k}$ with the SIP, there are $N = |\mathcal{H}|\left[(k-1)\binom{m}{k} + 1\right]$ vectors $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^m$ such that every matrix $\mathbf{A}$ with $L_{2k}(\mathbf{A}) > 0$ generates a dataset $Y = \{\mathbf{A}\mathbf{x}_1, \dots, \mathbf{A}\mathbf{x}_N\}$ with a stable k-sparse representation in $\mathbb{R}^m$ (with $\varepsilon(\delta_1, \delta_2)$ as in* (8)*).*

One can also easily verify that for every $k < m$, there are regular k-uniform hypergraphs with the SIP besides the obvious $\mathcal{H} = \binom{[m]}{k}$; for instance, take $\mathcal{H}$ to be the consecutive intervals of length k in some cyclic order on $[m]$. In this case, a direct consequence of Cor. 1 is rigorous verification of the lower bound $N = m(k-1)\binom{m}{k} + m$ for sufficient sample size from the introduction. Often, the SIP is achievable with fewer supports. For example, when $k = \sqrt{m}$, take $\mathcal{H}$ to be the $2k$ rows and columns formed by arranging $[m]$ into a square grid.

A practical implication of Thm. 1 is the following: there is an effective procedure sufficient to affirm if a proposed solution to Prob. 1 is indeed unique (up to noise and inherent ambiguities). One need simply to check that the matrix and codes satisfy the (computable) assumptions of Thm. 1 on $\mathbf{A}$ and the $\mathbf{x}_i$.

We furthermore note that unlike in previous works, the matrix $\mathbf{A}$ need not satisfy (2) to be recoverable from data. As an example for $k = 2$, let $\mathbf{A} = [\mathbf{e}_1, \dots, \mathbf{e}_5, \mathbf{v}]$ where $\mathbf{v} = \mathbf{e}_1 + \mathbf{e}_3 + \mathbf{e}_5$, and take $\mathcal{H}$ to be all consecutive pairs of $[m]$ arranged in cyclic order. Then, dictionary $\mathbf{A}$ satisfies the assumptions of Thm. 1 guaranteeing (6) without satisfying (2).

There are other less direct consequences of Thm. 1. For instance, we use it to prove uniqueness and stability of solutions to Prob. 2, the usual optimization problem of interest for those applying dictionary learning to their data.

Theorem 2. *If the assumptions of Thm. 1 hold, only now with more than $(k-1)\left[\binom{\overline{m}}{k} + |\mathcal{H}|k\binom{\overline{m}}{k-1}\right]$ vectors $\mathbf{x}_i$ supported in each $S \in \mathcal{H}$, then all solutions to Prob. 2 with $\eta \leq \varepsilon/2$ necessarily satisfy recovery inequalities* (6) *and* (7) *of Thm. 1.*

Another extension of Thm. 1 arises from the following analytic characterization of the spark condition. Let $\mathbf{A}$ be the $n \times m$ matrix of nm indeterminates A_{ij}. When real numbers are substituted for A_{ij}, the resulting matrix satisfies (2) if and only if the following polynomial is nonzero:

$$f(\mathbf{A}) := \prod_{S\in\binom{[m]}{2k}} \sum_{S'\in\binom{[n]}{2k}} (\det \mathbf{A}_{S',S})^2,$$

where for any $S' \in \binom{[n]}{2k}$ and $S \in \binom{[m]}{2k}$, the symbol $\mathbf{A}_{S',S}$ denotes the submatrix of entries A_{ij} with $(i,j) \in S' \times S$. We note that the large number of terms in this product is likely necessary due to the NP-hardness of deciding whether a given matrix $\mathbf{A}$ satisfies the spark condition [**?**].

Since f is analytic, having a single substitution of a real matrix $\mathbf{A}$ with $f(\mathbf{A}) \neq 0$ necessarily implies that the zeroes of f form a set of (Borel) measure zero. Fortunately, such a matrix $\mathbf{A}$ is easily constructed by adding rows of zeroes to any $\min(2k, m) \times m$ Vandermonde matrix as described above (so that each term in the product above for f is nonzero). Hence, almost every $n \times m$ matrix with $n \geq \min(2k, m)$ satisfies (2).

A similar phenomenon applies to datasets of vectors with a stable sparse representation. As in [**?**, Sec. IV], consider the "symbolic" dataset $Y = \{\mathbf{A}\mathbf{x}_1, \dots, \mathbf{A}\mathbf{x}_N\}$ generated by indeterminate $\mathbf{A}$ and indeterminate k-sparse $\mathbf{x}_1, \dots, \mathbf{x}_N$.

Theorem 3. *There is a polynomial g in the entries of $\mathbf{A} \in \mathbb{R}^{n\times m}$ and $\mathbf{x}_i \in \mathbb{R}^m$ with the following property: if g evaluates to a nonzero number and more than $(k-1)\binom{m}{k}$ of the resulting $\mathbf{x}_i$ are supported in each $S \in \mathcal{H}$ for some regular $\mathcal{H} \subseteq \binom{[m]}{k}$ with the SIP, then Y has a stable k-sparse representation in $\mathbb{R}^m$ (Def. 1). In particular, all – except for a Borel set of measure zero – substitutions impart to Y this property.*

Corollary 2. *Fix $m > k$, $n \geq \min(2k, m)$, and let the entries of $\mathbf{A} \in \mathbb{R}^{n\times m}$ and k-sparse $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^m$ be drawn independently from probability measures absolutely continuous with respect to the standard Borel measure. If more than $(k-1)\binom{m}{k}$ of the vectors $\mathbf{x}_i$ are supported in each $S \in \mathcal{H}$ for a regular $\mathcal{H} \subseteq \binom{[m]}{k}$ with the SIP, then Y has a stable k-sparse representation in $\mathbb{R}^m$ with probability one.*

Thus, drawing the dictionary and sparse codes from any continuous probability distribution almost certainly generates data with a stable sparse representation.

We remark that these results have a potential application to theoretical neuroscience by mathematically justifying one of the few hypothesized theories of bottleneck communication between sparsely active neural populations [**?**].

Proposition 1. *Sparse neural population activity is recoverable from noisy random linear compression by any method that solves Prob. 1 or 2; in particular, via biophysically plausible unsupervised sparse coding (e.g., [?], [?], [?]).*

We close this section with comments on optimality. Our linear scaling for ε in (8) is essentially optimal (e.g., see [**?**]), but a basic open problem remains: how many samples are necessary to determine the sparse coding model? If k is held fixed or if the size of the support set of reconstructing codes is known to be polynomial in $\overline{m}$ and k, then a practical (polynomial) amount of data suffices.[3] Reasons to be skeptical that this holds in general, however, can be found in [**?**], [**?**].

[3]In the latter case, a reexamination of the pigeonholing argument in the proof of Thm. 1 requires a polynomial number of samples distributed over a polynomial number of supports.

III. Proofs

We now begin our proof of Thm. 1 by showing how dictionary recovery (6) already implies sparse code recovery (7) when $\varepsilon < L_{2k}(\mathbf{A})/C_1$ (provided $\mathbf{A}$ satisfies (2)), temporarily assuming (without loss of generality) that $\overline{m} = m$. For $2k$-sparse $\mathbf{x} \in \mathbb{R}^m$, the triangle inequality gives $\|(\mathbf{A} - \mathbf{BPD})\mathbf{x}\|_2 \leq C_1\varepsilon\|\mathbf{x}\|_1 \leq C_1\varepsilon\sqrt{2k}\|\mathbf{x}\|_2$. Thus:

$$\begin{aligned}\|\mathbf{BPDx}\|_2 &\geq |\|\mathbf{Ax}\|_2 - \|(\mathbf{A} - \mathbf{BPD})\mathbf{x}\|_2| \\ &\geq \sqrt{2k}(L_{2k}(\mathbf{A}) - C_1\varepsilon)\|\mathbf{x}\|_2,\end{aligned}$$

since $\varepsilon < L_{2k}(\mathbf{A})/C_1$. Hence, $L_{2k}(\mathbf{BPD}) \geq L_{2k}(\mathbf{A}) - C_1\varepsilon > 0$, and (7) then follows from:

$$\begin{aligned}\|\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_i\|_1 &\leq \frac{\|\mathbf{BPD}(\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_i)\|_2}{L_{2k}(\mathbf{BPD})} \\ &\leq \frac{\|(\mathbf{BPD} - \mathbf{A})\mathbf{x}_i\|_2 + \|\mathbf{Ax}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2}{L_{2k}(\mathbf{BPD})} \\ &\leq \frac{\varepsilon(1 + C_1\|\mathbf{x}_i\|_1)}{L_{2k}(\mathbf{BPD})}.\end{aligned}$$

The heart of the matter is therefore (6), which we now establish, first in the important special case of $k = 1$.

Proof of Thm. 1 for $k = 1$. Since the only 1-uniform hypergraph with the SIP is $[m]$, we have $\mathbf{x}_i = c_i\mathbf{e}_i$ for $c_i \in \mathbb{R}\setminus\{\mathbf{0}\}$, $i \in [m]$. In this case, we require that $C_1 \geq 1/\min_{\ell\in[m]}|c_\ell|$.

Fix $\mathbf{A} \in \mathbb{R}^{n\times m}$ satisfying (2) and suppose that for some $\mathbf{B}$ and 1-sparse $\overline{\mathbf{x}}_i \in \mathbb{R}^{\overline{m}}$ we have $\|\mathbf{Ax}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \varepsilon < L_2(\mathbf{A})/C_1$ for all i. Then, there exist $\overline{c}_1, \ldots, \overline{c}_m \in \mathbb{R}$ and a map $\pi : [m] \to [\overline{m}]$ such that:

$$\|c_j\mathbf{A}_j - \overline{c}_j\mathbf{B}_{\pi(j)}\|_2 \leq \varepsilon, \quad \text{for } j \in [m]. \tag{9}$$

Note that $\overline{c}_j \neq 0$, since otherwise we have the contradiction $\|\mathbf{A}_j\|_2 \leq C_1|c_j|\|\mathbf{A}_j\|_2 < L_2(\mathbf{A}) \leq L_1(\mathbf{A}) = \min_{j\in[m]}\|\mathbf{A}_j\|_2$.

We now show that π is injective (in particular, a permutation if $\overline{m} = m$). Suppose that $\pi(i) = \pi(j) = \ell$ for some $i \neq j$ and ℓ. Then, $\|c_j\mathbf{A}_j - \overline{c}_j\mathbf{B}_\ell\|_2 \leq \varepsilon$ and $\|c_i\mathbf{A}_i - \overline{c}_i\mathbf{B}_\ell\|_2 \leq \varepsilon$. Scaling and summing these inequalities by $|\overline{c}_i|$ and $|\overline{c}_j|$, respectively, and applying the triangle inequality, we obtain:

$$\begin{aligned}(|\overline{c}_i| + |\overline{c}_j|)\varepsilon &\geq \|\mathbf{A}(\overline{c}_i c_j\mathbf{e}_j - \overline{c}_j c_i\mathbf{e}_i)\|_2 \\ &\geq (|\overline{c}_i| + |\overline{c}_j|)\, L_2(\mathbf{A}) \min_{\ell\in[m]}|c_\ell|,\end{aligned}$$

which contradicts the bound $\varepsilon < L_2(\mathbf{A})/C_1$. Hence, the map π is injective and therefore $\overline{m} \geq m$. Setting $\overline{J} = \pi([m])$ and letting $P = \left(\mathbf{e}_{\pi(1)}\cdots\mathbf{e}_{\pi(m)}\right)$ and $D = \operatorname{diag}(\frac{\overline{c}_1}{c_1}, \ldots, \frac{\overline{c}_m}{c_m})$, we see that (9) becomes, for all $j \in [m]$:

$$\|\mathbf{A}_j - \mathbf{B}_{\overline{J}}\mathbf{PD}_j\|_2 = \|\mathbf{A}_j - \frac{\overline{c}_j}{c_j}\mathbf{B}_{\pi(j)}\|_2 \leq \frac{\varepsilon}{|c_j|} \leq C_1\varepsilon.$$

□

We require a few additional tools to extend the proof to the general case $k < m$. These include a generalized notion of distance (Def. 3) and angle (Def. 4) between subspaces as well as a stability result in combinatorial matrix analysis (Lem. 1).

Definition 3. *For $\mathbf{u} \in \mathbb{R}^m$ and vector spaces $U, V \subseteq \mathbb{R}^m$, let* $\operatorname{dist}(\mathbf{u}, V) := \min\{\|\mathbf{u} - \mathbf{v}\|_2 : \mathbf{v} \in V\}$ *and define:*

$$d(U, V) := \max_{\mathbf{u}\in U,\ \|\mathbf{u}\|_2\leq 1} \operatorname{dist}(\mathbf{u}, V). \tag{10}$$

We note the following facts. If $U' \subseteq U$, then $d(U', V) \leq d(U, V)$ and [**?**, Ch. 4 Cor. 2.6]:

$$d(U, V) < 1 \implies \dim(U) \leq \dim(V). \tag{11}$$

Also, from [**?**, Lem. 3.2], we have:

$$\dim(U) = \dim(V) \implies d(U, V) = d(V, U). \tag{12}$$

Our result in combinatorial matrix analysis is the following.

Lemma 1. *Suppose the $n \times m$ matrix $\mathbf{A}$ has $L_{2\mathcal{H}}(\mathbf{A}) > 0$ for some r-regular $\mathcal{H} \subseteq \binom{[m]}{k}$ with the SIP. There exists $C_2 > 0$ for which the following holds for all $\varepsilon < L_2(\mathbf{A})/C_2$:*

If for some $n \times \overline{m}$ matrix $\mathbf{B}$ and map $\pi : \mathcal{H} \mapsto \binom{[\overline{m}]}{k}$,

$$d(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\pi(S)}) \leq \varepsilon, \quad \text{for } S \in \mathcal{H}; \tag{13}$$

then $\overline{m} \geq m$, and, provided $\overline{m} < mr/(r-1)$, there is a permutation matrix $\mathbf{P}$ and invertible diagonal $\mathbf{D}$ such that:

$$\|\mathbf{A}_j - \mathbf{B}_{\overline{J}}\mathbf{PD}_j\|_2 \leq C_2\varepsilon, \quad \text{for } j \in J, \tag{14}$$

for some nonempty $\overline{J} \subseteq [\overline{m}]$ and $J \subseteq [m]$ of size $\overline{m} - r(\overline{m} - m)$.

The constant $C_1 > 0$ in Thm. 1 is then given by[4]:

$$C_1(\mathbf{A}, \{\mathbf{x}_i\}_{i=1}^N, \mathcal{H}) := \frac{C_2(\mathbf{A}, \mathcal{H})}{\min_{S\in\mathcal{H}} L_k(\mathbf{AX}_{I(S)})}, \tag{15}$$

where, given vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$, we denote by $\mathbf{X}$ the $m \times N$ matrix with columns $\mathbf{x}_i$ and by $I(S)$ the set of indices i for which $\mathbf{x}_i$ is supported in S.

The constant $C_2 = C_2(\mathbf{A}, \mathcal{H})$, in turn, will be presented relative to a quantity from [**?**] used to analyze the convergence of the alternating projections algorithm. Specifically, in Lem. 3 below, the following definition is used to bound the distance between a point and the intersection of subspaces given an upper bound on its distance from each individual subspace.

Definition 4. *For a collection of real subspaces $\mathcal{V} = \{V_i\}_{i=1}^\ell$, define $\xi^2 := 0$ when $|\mathcal{V}| = 1$, and otherwise:*

$$\xi^2(\mathcal{V}) := 1 - \max \prod_{i=1}^{\ell-1} \sin^2\theta\left(V_i, \cap_{j>i} V_j\right), \tag{16}$$

*where the maximum is taken over all orderings of the V_i and the angle $\theta \in (0, \frac{\pi}{2}]$ is defined implicitly as [**?**, Def. 9.4]:*

$$\cos\theta(U, W) := \max\left\{|\langle \mathbf{u}, \mathbf{w}\rangle| : \begin{smallmatrix}\mathbf{u}\in U\cap(U\cap W)^\perp,\ \|\mathbf{u}\|_2\leq 1 \\ \mathbf{w}\in W\cap(U\cap W)^\perp,\ \|\mathbf{w}\|_2\leq 1\end{smallmatrix}\right\}.$$

[4]Note that $\|\mathbf{AX}_{I(S)}\mathbf{c}\|_2 \geq \sqrt{k}L_{\mathcal{H}}(\mathbf{A})\|\mathbf{X}_{I(S)}\mathbf{c}\|_2 \geq kL_{\mathcal{H}}(\mathbf{A})L_k(\mathbf{X}_{I(S)})\|\mathbf{c}\|_2$ for $S \in \mathcal{H}$ and k-sparse $\mathbf{c}$. Therefore, $L_k(\mathbf{AX}_{I(S)}) \geq \sqrt{k}L_{\mathcal{H}}(\mathbf{A})L_k(\mathbf{X}_{I(S)}) > 0$ since $L_{2\mathcal{H}}(\mathbf{A}) > 0$ and $L_k(\mathbf{X}_{I(S)}) > 0$ by general linear position of the $\mathbf{x}_i$. Thus, $C_1 > 0$.

Note that $\theta \in (0, \frac{\pi}{2}]$ implies $0 \leq \xi < 1$, and that $\xi(\mathcal{V}') \leq \xi(\mathcal{V})$ when $\mathcal{V}' \subseteq \mathcal{V}$.[5] The constant C_2 in Lem. 1 is then:

$$C_2(\mathbf{A}, \mathcal{H}) := \frac{(r+1)\max_{j\in[m]} \|\mathbf{A}_j\|_2}{1 - \max_{\mathcal{G}\in\binom{\mathcal{H}}{r+1}} \xi(\boldsymbol{\mathcal{A}}_{\mathcal{G}})}, \tag{17}$$

which we remark yields a constant C_1 consistent with what is required for the case $k = 1$ considered at the beginning of this section.[6]

The pragmatic reader should note that the explicit constants C_1 and C_2 are effectively computable: the quantity L_k may be calculated as the smallest singular value of a certain matrix, while the quantity ξ involves computing "canonical angles" between subspaces, which reduce again to an efficient singular value decomposition. There is no known fast computation of L_k in general, however, since even $L_k > 0$ is NP-hard [**?**]; although fixing k yields polynomial complexity. Moreover, calculating C_2 requires an exponential number of queries to ξ unless r is held fixed, too (e.g., the "cyclic order" hypergraphs described above have $r = k$). Thus, as presented, C_1 and C_2 are not efficiently computable.

Proof of Thm. 1 for $k < m$. We find a map $\pi : \mathcal{H} \to 2^{[\overline{m}]}$ for which the distance $d(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\pi(S)})$ is controlled by ε. Applying Lem. 1 then completes the proof.

Since there are more than $(k-1)\binom{\overline{m}}{k}$ vectors $\mathbf{x}_i$ supported in each $S \in \mathcal{H}$, the pigeonhole principle gives $\overline{S} \in \binom{[\overline{m}]}{k}$ and a set of k indices $K \subseteq I(S)$ with all $\overline{\mathbf{x}}_i$, $i \in K$, supported in $\pi(S) := \overline{S}$. It also follows from $L_{2\mathcal{H}}(\mathbf{A}) > 0$ and the general linear position of the $\mathbf{x}_i$ that $L_k(\mathbf{A}\mathbf{X}_K) > 0$; that is, the columns of the $n \times k$ matrix $\mathbf{A}\mathbf{X}_K$ form a basis for $\boldsymbol{\mathcal{A}}_S$.

Fixing $\mathbf{0} \neq \mathbf{y} \in \boldsymbol{\mathcal{A}}_S$, there then exists $\mathbf{0} \neq \mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^k$ such that $\mathbf{y} = \mathbf{A}\mathbf{X}_K\mathbf{c}$. Setting $\overline{\mathbf{y}} = \mathbf{B}\overline{\mathbf{X}}_K\mathbf{c} \in \boldsymbol{\mathcal{B}}_{\overline{S}}$, we have:

$$\begin{aligned}\|\mathbf{y} - \overline{\mathbf{y}}\|_2 = \|\sum_{i=1}^{k} c_i(\mathbf{A}\mathbf{X}_K - \mathbf{B}\overline{\mathbf{X}}_K)_i\|_2 &\leq \varepsilon \sum_{i=1}^{k} |c_i| \\ &\leq \varepsilon\sqrt{k}\|\mathbf{c}\|_2 \leq \frac{\varepsilon}{L_k(\mathbf{A}\mathbf{X}_K)}\|\mathbf{y}\|_2,\end{aligned}$$

where the last inequality follows directly from (5). From Def. 3:

$$d(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\overline{S}}) \leq \frac{\varepsilon}{L_k(\mathbf{A}\mathbf{X}_K)} \leq \varepsilon\frac{C_1}{C_2}, \tag{18}$$

where the second inequality is due to $L_k(\mathbf{A}\mathbf{X}_K) \geq L_k(\mathbf{A}\mathbf{X}_{I(S)})$ and (15). Finally, apply Lem. 1 with $\varepsilon < L_2(\mathbf{A})/C_1$. □

Proof of Thm. 2. We bound the number of k-sparse $\overline{\mathbf{x}}_i$ and then apply Thm. 1. Let n_p be the number of $\overline{\mathbf{x}}_i$ with $\|\overline{\mathbf{x}}_i\|_0 = p$. Since the $\mathbf{x}_i$ are all k-sparse, by (4) we have: $k\sum_{p=0}^{\overline{m}} n_p \geq \sum_{i=0}^{N} \|\mathbf{x}_i\|_0 \geq \sum_{i=0}^{N} \|\overline{\mathbf{x}}_i\|_0 = \sum_{p=0}^{\overline{m}} p n_p$. Hence,

$$\sum_{p=k+1}^{\overline{m}} n_p \leq \sum_{p=k+1}^{\overline{m}} (p-k)n_p \leq \sum_{p=0}^{k}(k-p)n_p \leq k\sum_{p=0}^{k-1} n_p, \tag{19}$$

demonstrating that the number of vectors $\overline{\mathbf{x}}_i$ that are *not* k-sparse is controlled by how many are $(k-1)$-sparse.

Next, observe that no more than $(k-1)|\mathcal{H}|$ of the $\overline{\mathbf{x}}_i$ share a support $\overline{S}$ of size less than k; otherwise, by the pigeonhole principle, at least k of these indices i belong to the same $K \subseteq I(S)$ for some $S \in \mathcal{H}$ and (as argued previously) (18) follows. Since the right-hand side of (18) is less than one, by (11) we have the contradiction $k = \dim(\boldsymbol{\mathcal{A}}_S) \leq \dim(\boldsymbol{\mathcal{B}}_{\overline{S}}) \leq |\overline{S}|$.

The total number of $(k-1)$-sparse vectors $\overline{\mathbf{x}}_i$ thus can not exceed $|\mathcal{H}|(k-1)\binom{\overline{m}}{k-1}$. By (19), no more than $|\mathcal{H}|k(k-1)\binom{\overline{m}}{k-1}$ vectors $\overline{\mathbf{x}}_i$ are not k-sparse. Since for every $S \in \mathcal{H}$ there are over $(k-1)\left[\binom{\overline{m}}{k} + |\mathcal{H}|k\binom{\overline{m}}{k-1}\right]$ vectors $\mathbf{x}_i$ supported there, it must be that more than $(k-1)\binom{\overline{m}}{k}$ of them have corresponding $\overline{\mathbf{x}}_i$ that are k-sparse. The result now follows from Thm. 1. □

Proof (sketch) of Thm. 3. Let $\mathbf{M}$ be the matrix with columns $\mathbf{A}\mathbf{x}_i$, $i \in [N]$. Consider the polynomial $\prod_{S\in\binom{[N]}{2k}} \sum_{S'\in\binom{[n]}{2k}} (\det \mathbf{M}_{S',S})^2$ in the indeterminate entries of $\mathbf{A}$ and $\mathbf{x}_i$, with notation as in Sec. II. It can be checked that when this polynomial is nonzero for a substitution of real numbers for the indeterminates, all of the genericity requirements on $\mathbf{A}$ and $\mathbf{x}_i$ in our proofs of stability in Thm. 1 are satisfied (in particular, the spark condition (2) on $\mathbf{A}$). □

Proof (sketch) of Cor. 2. First, note that if a set of measure spaces $\{(X_\ell, \Sigma_\ell, \nu_\ell)\}_{\ell=1}^{p}$ has that ν_ℓ is absolutely continuous with respect to μ for all $\ell \in [p]$, where μ is the standard Borel measure on $\mathbb{R}$, then the product measure $\prod_{\ell=1}^{p} \nu_\ell$ is absolutely continuous with respect to the standard Borel product measure on $\mathbb{R}^p$. By Thm. 3, there is a polynomial that is nonzero when Y has a stable k-sparse representation in $\mathbb{R}^m$; in particular, stability holds almost surely. □

IV. Discussion

A motivation for this work was the emergence of seemingly characteristic representations from sparse coding models fit to natural data, despite the varied assumptions underlying the many algorithms in current use. To this end, we have taken an important step toward unifying the great number of publications on the topic by demonstrating very general, deterministic conditions under which identification of parameters in this model is not only possible but also robust to the inevitable uncertainty permeating measurement and model choice.

We have shown that, given sufficient data, the problem of seeking a dictionary and sparse codes with minimal average support size (Prob. 2) reduces to an instance of Prob. 1, to which our main result (Thm. 1) applies: every dictionary and sequence of sparse codes consistent with the data are equivalent up to inherent relabeling/scaling ambiguities and a

[5] We acknowledge the counter-intuitive property that $\theta = \pi/2$ when $U \subseteq W$.

[6] For $\mathbf{x}_i = c_i\mathbf{e}_i$, the denominator in (15) becomes $\min_{i\in[m]} |c_i|\|\mathbf{A}_i\|_2$, hence with $r = 1$ we have $C_1 \geq 2/\min_{i\in[m]} |c_i|$.

discrepancy (error) that scales linearly with the measurement noise or modeling inaccuracy. The constants we provide are explicit and computable; as such, there is an effective procedure that sufficiently affirms if a proposed solution to Probs. 1 or 2 is indeed unique up to noise and inherent ambiguities.

An immediate application of our theoretical work is Prop. 1, which certifies the validity of a theory of biologically-plausible bottleneck communication in the brain: that sparse coding in a compressed space of neural activity can recover sparse codes sent through a randomly-constructed (but unknown) noisy wiring bottleneck [**?**].[7]

Beyond an original extension of existing noiseless guarantees [**?**] to the noisy regime and their application to Prob. 2, a major innovation in our work is a theory of combinatorial designs for support sets key to the identification of the dictionary. We incorporate this idea into a fundamental lemma in matrix theory (Lem. 1) that draws upon the definition of a new matrix lower bound induced by a hypergraph. Insights enabled by our combinatorial approach include: 1) a subset of dictionary elements is recoverable even if dictionary size is overestimated, 2) data require only a polynomial number of distinct sparse supports, and 3) the spark condition is not a necessary property of recoverable dictionaries.

The absence of any assumption at all about dictionaries that solve Prob. 1 was a major technical difficulty in proving Thm. 1. We sought such a general guarantee because of the practical difficulty of ensuring that an algorithm maintain a dictionary satisfying the spark condition (2) at each iteration, an implicit requirement of all previous works except [**?**]; indeed, even certifying a dictionary has this property is NP-hard [**?**].

In fact, uniqueness guarantees with minimal assumptions apply to all areas of data science and engineering that utilize learned sparse structure. For example, several groups have applied compressed sensing to signal processing tasks: MRI analysis [**?**], image compression [**?**], and, more recently, the design of an ultrafast camera [**?**]. Given such effective uses of compressed sensing, it is only a matter of time before these systems incorporate dictionary learning to encode and decode signals (e.g., in a device that learns structure from motion [**?**]), just as scientists have used it to make sense of their data [**?**], [**?**], [**?**], [**?**]. Assurances such as those offered by our theorems certify that different devices (with different initialization, etc.) will learn equivalent representations given enough data from statistically identical systems.[8] Indeed, it seems a main reason for the sustained interest in dictionary learning as an unsupervised method for data analysis is the assumed well-posedness of parameter identification in the sparse coding model, confirmation of which forms the core of our theoretical findings.

Acknowledgement. We thank F. Sommer and D. Rhea for early thoughts, and I. Morris for posting (12) online.

[7]We refer the reader to [**?**] for more on interactions between dictionary learning and neuroscience.

[8]To contrast with the current hot topic of "Deep Learning", there are few such uniqueness guarantees for these models of data; moreover, even small noise can dramatically alter their output [**?**].

V. Appendix

In this section, we prove Lem. 1 and some auxiliary lemmas.

Lemma 2. *If $f : V \to W$ is an injective function, then $f\left(\cap_{i=1}^{\ell} V_i\right) = \cap_{i=1}^{\ell} f\left(V_i\right)$ for any $V_1, \ldots, V_\ell \subseteq V$. ($f(\emptyset) := \emptyset$.)*

Proof. By induction, it is enough to prove the trivial case $\ell = 2$. □

In particular, if a matrix $\mathbf{A}$ satisfies $L_{2\mathcal{H}}(\mathbf{A}) > 0$, then letting V be the set of vectors with supports in $2\mathcal{H}$, we have $\boldsymbol{\mathcal{A}}_{\cap\mathcal{G}} = \cap\boldsymbol{\mathcal{A}}_{\mathcal{G}}$ for all $\mathcal{G} \subseteq \mathcal{H}$.

Lemma 3. *Let $\mathcal{V} = \{V_i\}_{i=1}^k$ be a set of two or more subspaces of $\mathbb{R}^m$ and let $V = \cap_{i=1}^k V_i$. For $\mathbf{x} \in \mathbb{R}^m$, we have (recall Defs. 3 & 4):*

$$\operatorname{dist}(\mathbf{x}, V) \leq \frac{1}{1 - \xi(\mathcal{V})} \sum_{i=1}^{k} \operatorname{dist}(\mathbf{x}, V_i). \tag{20}$$

Proof. Recall the projection onto the subspace $V \subseteq \mathbb{R}^m$ is the mapping $\Pi_V : \mathbb{R}^m \to V$ that associates with each $\mathbf{x}$ its unique nearest point in V; i.e., $\|\mathbf{x} - \Pi_V \mathbf{x}\|_2 = \operatorname{dist}(\mathbf{x}, V)$. Next, observe:

$$\begin{aligned}\|\mathbf{x} - \Pi_V \mathbf{x}\|_2 &\leq \|\mathbf{x} - \Pi_{V_k}\mathbf{x}\|_2 + \|\Pi_{V_k}\mathbf{x} - \Pi_{V_k}\Pi_{V_{k-1}}\mathbf{x}\|_2 \\ &\quad + \cdots + \|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\mathbf{x} - \Pi_V\mathbf{x}\|_2 \\ &\leq \|\Pi_{V_k}\cdots\Pi_{V_1}\mathbf{x} - \Pi_V\mathbf{x}\|_2 + \sum_{\ell=1}^{k}\|\mathbf{x} - \Pi_{V_\ell}\mathbf{x}\|_2,\end{aligned} \tag{21}$$

using the triangle inequality and that the spectral norm satisfies $\|\Pi_{V_\ell}\|_2 \leq 1$ for all ℓ (since Π_{V_ℓ} are orthogonal projections).

The desired result, (20), now follows by bounding the second term on the right-hand side using the following fact [**?**, Thm. 9.33]:

$$\|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\mathbf{x} - \Pi_V\mathbf{x}\|_2 \leq z\|\mathbf{x}\|_2, \tag{22}$$

for $z^2 = 1 - \prod_{\ell=1}^{k-1}(1 - z_\ell^2)$ and $z_\ell = \cos\theta\left(V_\ell, \cap_{s=\ell+1}^{k} V_s\right)$, recalling θ from Def. 4. Together with $\Pi_{V_\ell}\Pi_V = \Pi_V$ for all $\ell \in [k]$ and $\Pi_V^2 = \Pi_V$, this implies that $\|\Pi_{V_k}\cdots\Pi_{V_1}\mathbf{x} - \Pi_V\mathbf{x}\|_2$ satisfies: $\|(\Pi_{V_k}\cdots\Pi_{V_1} - \Pi_V)(\mathbf{x} - \Pi_V\mathbf{x})\|_2 \leq z\|\mathbf{x} - \Pi_V\mathbf{x}\|_2$.

Finally, substituting this into (21) and rearranging produces (20) after substituting $\xi(\mathcal{V})$ for z. □

Lemma 4. *Fix an r-regular hypergraph $\mathcal{H} \subseteq 2^{[m]}$ satisfying the SIP. If the map $\pi : \mathcal{H} \to 2^{[\overline{m}]}$ has $\sum_{S\in\mathcal{H}}|\pi(S)| \geq \sum_{S\in\mathcal{H}}|S|$ and:*

$$|\cap\,\pi(\mathcal{G})| \leq |\cap\,\mathcal{G}|, \quad \text{for } \mathcal{G} \in \binom{\mathcal{H}}{r} \cup \binom{\mathcal{H}}{r+1}, \tag{23}$$

then $\overline{m} \geq m$, and if $(r-1)\overline{m} < mr$ then the map $i \mapsto \cap\pi(\sigma(i))$ is an injective map to $[\overline{m}]$ from some $J \subseteq [m]$ of size $\overline{m} - r(\overline{m} - m)$.

Proof. Consider the set: $T_1 := \{(i, S) : i \in \pi(S), S \in \mathcal{H}\}$, which numbers $|T_1| = \sum_{S\in\mathcal{H}}|\pi(S)| \geq \sum_{S\in\mathcal{H}}|S| = \sum_{i\in[m]}\deg_{\mathcal{H}}(i) = mr$. Note that $|T_1| \leq \overline{m}r$ by (23), since otherwise pigeonholing the elements of T_1 with respect to their

possible first indices $[\overline{m}]$ would imply $\deg_{\mathcal{H}}(i) \geq r$ for some $i \in [m]$. Thus, $\overline{m} \geq m$.

Suppose $\overline{m}(r-1) < mr$, so that $|T_1| \geq mr = \overline{m}(r-1) + p$ for positive $p := mr - \overline{m}(r-1)$. Pigeonholing T_1 into $[\overline{m}]$ again, there must be at least r elements of T_1 that share a first index. Let $i_1 \in [\overline{m}]$ be such an index and note that, in fact, no more than r elements of T_1 can have i_1 as a first index; otherwise, the set $\mathcal{G}_1$ of their second indices would contain $r+1$ elements with a non-empty intersection by (23), contradicting r-regularity of $\mathcal{H}$. Thus, $|\mathcal{G}_1| = r$, and (23) implies $|\cap \mathcal{G}_1| = 1$ by r-regularity and the SIP. If $p = 1$, then we are done. Otherwise, define $T_2 := T_1 \setminus \{(i,S) \in T_1 : i = i_1\}$, which contains $|T_2| = |T_1| - r \geq (\overline{m}-1)(r-1) + (p-1)$ ordered pairs having $\overline{m}-1$ distinct first indices. Repeating the above arguments thus produces a distinct $i_2 \in [\overline{m}]$ and $\mathcal{G}_2 \in \binom{\mathcal{H}}{r}$ with (necessarily, by r-regularity and the SIP) distinct singleton $\cap\mathcal{G}_2 \in [m]$. Finally, iterating this p times in total yields the set of singletons $J = \{\cap\mathcal{G}_1, \ldots, \cap\mathcal{G}_p\}$. □

Proof of Lem. 1. We begin by showing that $\dim(\boldsymbol{\mathcal{B}}_{\pi(S)}) = \dim(\boldsymbol{\mathcal{A}}_S)$ for all $S \in \mathcal{H}$. Note that since $\|\mathbf{A}\mathbf{x}\|_2 \leq \max_j \|\mathbf{A}_j\|_2 \|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_1 \leq \sqrt{k}\|\mathbf{x}\|_2$ for all k-sparse $\mathbf{x}$, by (5) we have $L_2(\mathbf{A}) \leq \max_j \|\mathbf{A}_j\|_2$ and therefore (as $\xi > 0$), the right-hand side of (13) is less than one. From (11), we have $|\pi(S)| \geq \dim(\boldsymbol{\mathcal{B}}_{\pi(S)}) \geq \dim(\boldsymbol{\mathcal{A}}_S) = |S|$, with the equality by injectivity of $\mathbf{A}_S$. As $|\pi(S)| = |S|$, the claim follows. Note, therefore, that $\mathbf{B}_{\pi(S)}$ has full-column rank for all $S \in \mathcal{H}$.

We next verify that (23) holds. Fixing $\mathcal{G} \in \binom{\mathcal{H}}{r} \cup \binom{\mathcal{H}}{r+1}$, it suffices to show that $d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) < 1$, since by (11) we then have $|\cap\,\pi(\mathcal{G})| = \dim(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}) \leq \dim(\boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) = |\cap\,\mathcal{G}|$, with equalities by the full column-ranks of $\mathbf{A}_S$ and $\mathbf{B}_{\pi(S)}$ for all $S \in \mathcal{H}$.[9] Observe that $d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) \leq d\left(\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})}, \cap\boldsymbol{\mathcal{A}}_{\mathcal{G}}\right)$, since $d(U', V) \leq d(U, V)$ whenever $U' \subseteq U$ and $\boldsymbol{\mathcal{A}}_{\cap\mathcal{G}} = \cap\boldsymbol{\mathcal{A}}_{\mathcal{G}}$ by Lem. 2. Recalling Def. 3, applying Lem. 3 and carrying the supremum through the sum yields:

$$d\left(\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})}, \cap\boldsymbol{\mathcal{A}}_{\mathcal{G}}\right) \leq \sum_{S\in\mathcal{G}} \frac{d\left(\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_S\right)}{1 - \xi(\boldsymbol{\mathcal{A}}_{\mathcal{G}})}. \tag{24}$$

Since $\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})} \subseteq \boldsymbol{\mathcal{B}}_{\pi(S)}$ for all $S \in \mathcal{G}$, the numerator of each term above is bounded by $d\left(\boldsymbol{\mathcal{B}}_{\pi(S)}, \boldsymbol{\mathcal{A}}_S\right)$ and, recalling (12) (since $\dim(\boldsymbol{\mathcal{B}}_{\pi(S)}) = \dim(\boldsymbol{\mathcal{A}}_S)$), in turn by $d\left(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\pi(S)}\right) \leq \varepsilon$. Thus:

$$d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) \leq \frac{|\mathcal{G}|\varepsilon}{1 - \xi(\boldsymbol{\mathcal{A}}_{\mathcal{G}})} \leq \frac{C_2\varepsilon}{\max_j \|\mathbf{A}_j\|_2}, \tag{25}$$

which is less than one, since $C_2\varepsilon < L_2(\mathbf{A}) \leq \max_j \|\mathbf{A}_j\|_2$.

By Lem. 4, the association $i \mapsto \cap\pi(\sigma(i))$ is an injective map $\overline{\pi} : J \to [\overline{m}]$ for some $J \subseteq [m]$ of size $\overline{m} - r(\overline{m} - m)$, and $\mathbf{B}_{\overline{\pi}(i)} \neq \mathbf{0}$ for all $i \in J$ since the columns of $\mathbf{B}_{\pi(S)}$ are linearly independent for all $S \in \mathcal{H}$. Letting $\overline{\varepsilon} := C_2\varepsilon / \max_i \|\mathbf{A}_i\|_2$, it follows from (12) and (25) that $d\left(\boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{B}}_{\overline{\pi}(i)}\right) \leq \overline{\varepsilon}$ for all $i \in J$. Setting $c_i := \|\mathbf{A}_i\|_2^{-1}$ so that $\|c_i\mathbf{A}\mathbf{e}_i\|_2 = 1$, by Def. 3 we have for all $i \in J$:

$$\min_{\overline{c}_i \in \mathbb{R}} \|c_i\mathbf{A}\mathbf{e}_i - \overline{c}_i\mathbf{B}\mathbf{e}_{\overline{\pi}(i)}\|_2 \leq d\left(\boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{B}}_{\overline{\pi}(i)}\right) \leq \overline{\varepsilon},$$

for $\overline{\varepsilon} < L_2(\mathbf{A}) \min_{i\in[m]} |c_i|$. But this is exactly the supposition in (9), with J and $\overline{\varepsilon}$ in place of $[m]$ and ε, respectively. Applying the arguments of the case $k = 1$ in Sec. III to the submatrix $\mathbf{A}_J$ then finally yields $\|\mathbf{A}_j - \mathbf{B}_{\overline{J}}\mathbf{P}\mathbf{D}_j\|_2 \leq \overline{\varepsilon}/|c_j| \leq C_2\varepsilon$ for $j \in J$. □

[9] Note that if ever $\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})} \neq \mathbf{0}$ while $\cap\mathcal{G} = \emptyset$, we would have $d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \mathbf{0}) = 1$. This cannot be the case, however, since the deduction that follows would then lead to a contradiction.