

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

Charles J. Garfinkle<sup>a,b,1</sup> and Christopher J. Hillar<sup>a,1</sup>

<sup>a</sup>Redwood Center for Theoretical Neuroscience, Berkeley, CA, USA; <sup>b</sup>Helen Wills Neuroscience Institute, UC Berkeley

**Dictionary learning for sparse linear coding has exposed characteristic properties of natural signals. However, a universal theorem guaranteeing the consistency of estimation in this model is lacking. Here, we prove that for all diverse enough datasets generated from the sparse coding model, latent dictionaries and codes are uniquely and stably determined up to measurement error. Applications are given to data analysis, engineering, and neuroscience.**

Sparse coding, dictionary learning, matrix factorization, compressed sensing, inverse problems, blind source separation

**B**lind source separation is a classical problem in signal processing (1). In one common modern formulation, each of  $N$  observed  $n$ -dimensional signals 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 (2) for a comprehensive review of this and related models). Approximating solutions to this sparsity-constrained inverse problem have provided insight into the structure of many signal classes lacking domain-specific formal models (e.g., in vision (3)). In particular, it has been shown that response properties of simple-cell neurons in mammalian visual cortex emerge from optimizing a dictionary to represent small patches of natural images (4–7), a major advance in computational neuroscience. A curious aspect of this finding is that the latent waveforms (e.g., “Gabor” wavelets) estimated from data appear to be canonical (8); i.e., they are found in learned dictionaries independent of algorithm or image training set.

Motivated by these discoveries and earlier work in the theory of neural communication (9), we address when dictionaries and the sparse representations they induce are uniquely determined by data. Answers to this question also have other real-world implications. For example, a sparse coding analysis of local painting style can be used for forgery detection (10, 11), but only if all dictionaries consistent with training data do not differ appreciably in their ability to sparsely encode new samples. Fortunately, algorithms with proven recovery of generating dictionaries under certain conditions have recently been proposed (see (12, Sec. I–E) for a summary of the state-of-the-art). Few theorems, however, can be cited to explain this uniqueness more broadly; in particular, a universal guarantee in the context of noise has yet to emerge.

Here, we prove very generally that uniqueness and stability is an expected property of the sparse linear coding model. More specifically, dictionaries that preserve sparse codes (e.g., satisfy a “spark condition”) 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). In fact, provided  $n \geq \min(2k, m)$ , in almost all cases the dictionary learning problem is well-posed (as per Hadamard (13)) given enough data (Cor. 2). Moreover, these explicit, algorithm-independent guarantees hold without assuming the recovered matrix satisfies a spark condition, and even when

the number  $m$  of dictionary elements is unknown.

More formally, let  $\mathbf{A} \in \mathbb{R}^{n \times m}$  be a matrix with columns  $\mathbf{A}_j$  ( $j = 1, \dots, m$ ) and let dataset  $Z$  consist of measurements:

$$\mathbf{z}_i = \mathbf{A}\mathbf{x}_i + \mathbf{n}_i, \quad i = 1, \dots, N, \quad [1]$$

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  $\mathbf{A}\mathbf{x}_i$ . The first mathematical problem we address is the following.

**Problem 1** (Sparse linear coding). *Find a matrix  $\mathbf{B}$  and  $k$ -sparse  $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_N \in \mathbb{R}^m$  so that  $\|\mathbf{z}_i - \mathbf{B}\bar{\mathbf{x}}_i\|_2 \leq \eta$  for all  $i$ .*

Note that any particular solution  $(\mathbf{B}, \bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_N)$  to this problem gives rise to an orbit of equivalent solutions  $(\mathbf{B}\mathbf{P}, \mathbf{D}^{-1}\mathbf{P}^\top \bar{\mathbf{x}}_1, \dots, \mathbf{D}^{-1}\mathbf{P}^\top \bar{\mathbf{x}}_N)$ , where  $\mathbf{P}$  is any permutation matrix and  $\mathbf{D}$  any invertible diagonal. Previous theoretical work addressing the noiseless case  $\eta = 0$  (e.g., (14–17)) has shown that a solution to Prob. 1 (when it exists) is indeed unique up to this inherent ambiguity provided the  $\mathbf{x}_i$  are sufficiently diverse and the generating matrix  $\mathbf{A}$  satisfies the *spark condition* from compressed sensing:

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

which would be, in any case, necessary for uniqueness of the  $\mathbf{x}_i$ . Here, we study stability in the practical setting of 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  $\delta_1$  and  $\delta_2$  such that if  $\mathbf{B}$  and  $k$ -sparse  $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_N \in \mathbb{R}^m$  satisfy:*

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

## Significance Statement

Many natural signals (visual scenery, speech, EEG, etc.) lacking domain-specific formal models can nonetheless be usefully characterized as linear combinations of few elementary waveforms drawn from a large “dictionary”. We give general conditions guaranteeing when such dictionaries are uniquely and stably determined by data. The result justifies the use of this model as a constraint for blind source separation and helps explain the observed universality of emergent representations in sparse models of certain natural phenomena.

The authors declare no conflict of interest.

<sup>1</sup>To whom correspondence should be addressed. E-mail: chaz@berkeley.edu, chillar@msri.org

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

$$\|\mathbf{A}_j - \mathbf{BPD}_j\|_2 \leq \delta_1 \quad \text{and} \quad \|\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top \bar{\mathbf{x}}_i\|_1 \leq \delta_2. \quad [3]$$

To see how Def. 1 directly relates to Prob. 1, suppose that  $Y$  has a stable  $k$ -sparse representation in  $\mathbb{R}^m$  and fix  $\delta_1, \delta_2$  to be the desired recovery accuracy in [3]. Consider now any dataset  $Z$  generated as in [1] with  $\eta \leq \frac{1}{2}\varepsilon(\delta_1, \delta_2)$ . Then from the triangle inequality, it follows that any matrix  $\mathbf{B}$  and  $k$ -sparse  $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_N \in \mathbb{R}^m$  solving Prob. 1 are necessarily within  $\delta_1, \delta_2$  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. 2 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 that extends our guarantees to the following more common optimization formulation of the dictionary learning problem (Thm. 2).

**Problem 2.** Find  $\mathbf{B}$  and  $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_N \in \mathbb{R}^m$  that solve:

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

We then sketch the proof of a probabilistic extension of Thm. 1 to random data and dictionaries (Thm. 3 and Cor. 2). Finally, in Sec. 3, we discuss both theoretical and practical applications of our main mathematical findings.

## 1. 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 version of the spark condition. The lower bound of a matrix  $\mathbf{M} \in \mathbb{R}^{n \times m}$  is the largest  $\alpha$  such that  $\|\mathbf{M}\mathbf{x}\|_2 \geq \alpha\|\mathbf{x}\|_2$  for all  $\mathbf{x} \in \mathbb{R}^m$  (18). By compactness of the unit sphere, every injective linear map has 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 the lower bound to a domain-restricted union of subspaces model (19) derived from a hypergraph  $\mathcal{H} \subseteq \binom{[m]}{k}$ . Letting  $\mathbf{M}_J$  denote the submatrix formed by the columns of  $\mathbf{M}$  indexed by  $J \subseteq [m]$  (with  $\mathbf{M}_\emptyset := \mathbf{0}$ ), we 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  $\mathcal{H} = \binom{[m]}{k}$ .<sup>\*</sup> 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. The following is a precise statement of our main result. We leave the quantity  $C_1 = C_1(\mathbf{A}, \{\mathbf{x}_i\}_{i=1}^N, \mathcal{H})$  undefined until Eq. [15].

**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<sup>†</sup>  $\varepsilon < L_2(\mathbf{A})/C_1$ :

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

$$\|\mathbf{A}_j - \mathbf{B}_J \mathbf{P} \mathbf{D}_J\|_2 \leq C_1 \varepsilon, \quad \text{for all } j \in J, \quad [6]$$

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

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

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

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

In words, Thm. 1 says that the smaller the difference  $\bar{m} - m$  between recovery and original latent dictionary sizes, 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 dictionary  $\mathbf{B}$  and codes  $\bar{\mathbf{x}}_i$ . In the particular case when  $\bar{m} = m$ , the theorem 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  $\varepsilon$  in Def. 1 given by:

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

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  $\gamma_i$ ). Thus, the assumptions of Thm. 1 are easily met, leading to the following direct application to uniqueness in sparse linear coding.

**Corollary 1.** Given a regular hypergraph  $\mathcal{H} \subseteq \binom{[m]}{k}$  with the SIP, there are  $N = |\mathcal{H}|[(k-1)\binom{m}{k} + 1]$  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]).

<sup>\*</sup>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  $\ell_2$ -norm columns (20).

<sup>†</sup>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  $\mathbf{B} = [\mathbf{0}, \frac{1}{2}(\mathbf{e}_1 + \mathbf{e}_2), \mathbf{e}_3, \dots, \mathbf{e}_m]$  and 1-sparse codes  $\bar{\mathbf{x}}_i = \mathbf{e}_2$  for  $i = 1, 2$  and  $\bar{\mathbf{x}}_i = \mathbf{e}_i$  for  $i \geq 3$  satisfy  $\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\bar{\mathbf{x}}_i\|_2 \leq \varepsilon$  but nonetheless violate [6].

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. In many cases, however, the SIP is achievable with far fewer supports. For example, when  $k = \sqrt{m}$ , simply take  $\mathcal{H}$  to be the  $2k$  rows and columns formed by arranging  $[m]$  in a square grid.

Another 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). Simply 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 (14–17), the matrix  $\mathbf{A}$  need not satisfy the spark condition to be recoverable from data. As a simple 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 underlying [6] without satisfying [2].

There are other less direct consequences of Thm. 1. For instance, we use it to prove uniqueness for the optimization formulation of sparse linear coding, Prob. 2, the main object of interest for those applying dictionary learning to their data.

**Theorem 2.** *If the assumptions of Thm. 1 hold, only now with more than  $(k-1) \left[ \binom{\bar{m}}{k} + |\mathcal{H}|k \binom{\bar{m}}{k-1} \right]$  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 (21).

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 (17, 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  $\mathbf{A}$  and  $\mathbf{x}_i$  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, choosing the dictionary and codes “randomly” almost certainly generates data with a stable sparse representation.

We remark that these results have an important application to theoretical neuroscience by mathematically justifying one of the few hypothesized theories of bottleneck communication between sparsely active neural populations (9).

**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 linear coding (e.g., (22, 23)).*

We close this section with comments on optimality. Our linear scaling for  $\varepsilon$  in [8] is essentially optimal (e.g., see (24)), but a basic open problem remains: how many samples are necessary to determine the sparse linear coding model? If  $k$  is held fixed or if the size of the support set of reconstructing codes is known to be polynomial in  $\bar{m}$  and  $k$ , then a practical (polynomial) amount of data suffices.<sup>‡</sup> Reasons to be skeptical that this holds in general, however, can be found in (21, 25).

## 2. 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  $\bar{m} = m$ . 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:

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

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:

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

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  $c_i \in \mathbb{R} \setminus \{0\}$ ,  $i \in [m]$ . In this case, we may take any  $C_1 \geq 1/\min_{i \in [m]} |c_i|$ .

Fix  $\mathbf{A} \in \mathbb{R}^{n \times m}$  satisfying [2] and suppose that for some  $\mathbf{B}$  and 1-sparse  $\bar{\mathbf{x}}_i \in \mathbb{R}^{\bar{m}}$  we have  $\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\bar{\mathbf{x}}_i\|_2 \leq \varepsilon < L_2(\mathbf{A})/C_1$

<sup>‡</sup>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.

for all  $i$ . Then, there exist  $\bar{c}_1, \dots, \bar{c}_m \in \mathbb{R}$  and a map  $\pi : [m] \rightarrow [\bar{m}]$  such that:

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

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

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

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

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

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

□

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  $\text{dist}(\mathbf{u}, V) := \inf\{\|\mathbf{u} - \mathbf{v}\|_2 : \mathbf{v} \in V\}$  and define:

$$d(U, V) := \sup_{\mathbf{u} \in U, \|\mathbf{u}\|_2 \leq 1} \text{dist}(\mathbf{u}, V). \quad [10]$$

We note the following facts. If  $U' \subseteq U$ , then  $d(U', V) \leq d(U, V)$  and (26, Ch. 4 Cor. 2.6):

$$d(U, V) < 1 \implies \dim(U) \leq \dim(V). \quad [11]$$

Also, from (27, Lem. 3.2), we have:

$$\dim(U) = \dim(V) \implies d(U, V) = d(V, U). \quad [12]$$

Our result in combinatorial matrix analysis is the following. Let  $\mathcal{M}_S$  denote the column-span of a submatrix  $\mathbf{M}_S$ .

**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  $n \times \bar{m}$  matrix  $\mathbf{B}$  and map  $\pi : \mathcal{H} \mapsto \binom{[\bar{m}]}{k}$ , we have:

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

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

$$\|\mathbf{A}_j - \mathbf{B}_{\bar{J}} \mathbf{P} D_j\|_2 \leq C_2 \varepsilon, \quad \text{for } j \in J, \quad [14]$$

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

The constant  $C_1 > 0$  in Thm. 1 is then given by<sup>§</sup>:

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

where, given vectors  $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^m$ , 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 (28) used to analyze the convergence of the alternating projections algorithm. More 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 subspaces  $V_1, \dots, V_\ell \subseteq \mathbb{R}^m$ , set  $\xi := 1$  when  $\ell = 1$  and define for  $\ell \geq 2$ :

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

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 (28, Def. 9.4):

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

Note that  $\theta \in (0, \frac{\pi}{2}]$  implies  $0 < \xi \leq 1$ .<sup>¶</sup> The constant  $C_2$  in Lem. 1 can then be expressed as:

$$C_2(\mathbf{A}, \mathcal{H}) := \frac{(r+1) \max_{j \in [m]} \|\mathbf{A}_j\|_2}{\min_{\mathcal{G} \subseteq \binom{[n]}{r} \cup \binom{[n]}{r+1}} \xi(\{\mathcal{A}_S\}_{S \in \mathcal{G}})}, \quad [16]$$

which we remark is consistent with the assumption on  $C_1$  in the proof of the case  $k = 1$  at the beginning of this section.

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  $\xi$  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 (21); although fixing  $k$  yields polynomial complexity. Moreover, calculating  $C_2$  requires an exponential number of queries to  $\xi$ . Thus, as presented,  $C_1$  and  $C_2$  are not efficiently computable.

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

Since there are more than  $(k-1)\binom{[\bar{m}]}{k}$  vectors  $\mathbf{x}_i$  supported in  $S$ , the pigeonhole principle gives  $\bar{S} \in \binom{[\bar{m}]}{k}$  and a set of  $k$  indices  $K \subseteq I(S)$  with all  $\bar{\mathbf{x}}_i$ ,  $i \in K$ , supported in  $\bar{S}$ .

It follows from  $L_{2\mathcal{H}}(\mathbf{A}) > 0$  and the general linear position of the  $\mathbf{x}_i$  that  $L(\mathbf{A} \mathbf{X}_K) > 0$ ; that is, the columns of the  $n \times k$  matrix  $\mathbf{A} \mathbf{X}_K$  form a basis for  $\mathcal{A}_S$ . Fixing  $\mathbf{0} \neq \mathbf{y} \in \mathcal{A}_S$ , there

<sup>§</sup>Note that  $\|\mathbf{A} \mathbf{X}_{I(S)} \mathbf{c}\|_2 \geq \sqrt{k} L_{\mathcal{H}}(\mathbf{A}) \|\mathbf{X}_{I(S)} \mathbf{c}\|_2 \geq k L_{\mathcal{H}}(\mathbf{A}) L_k(\mathbf{X}_{I(S)}) \|\mathbf{c}\|_2$  for  $S \in \mathcal{H}$  and  $k$ -sparse  $\mathbf{c}$ . Therefore,  $L_k(\mathbf{A} \mathbf{X}_{I(S)}) \geq \sqrt{k} L_{\mathcal{H}}(\mathbf{A}) L_k(\mathbf{X}_{I(S)}) > 0$  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,  $C_1 > 0$ .

<sup>¶</sup>We acknowledge the counter-intuitive property that  $\theta = \pi/2$  when  $U \subseteq W$ .



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

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

where the last inequality follows directly from [5]. From Def. 3:

$$d(\mathcal{A}_S, \mathcal{B}_{\bar{S}}) \leq \frac{\varepsilon}{L(\mathbf{A}\mathbf{X}_K)} \leq \varepsilon \frac{C_1}{C_2}, \quad [17]$$

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

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

$$\sum_{p=k+1}^{\bar{m}} n_p \leq \sum_{p=k+1}^{\bar{m}} (p-k)n_p \leq \sum_{p=0}^k (k-p)n_p \leq k \sum_{p=0}^{k-1} n_p, \quad [18]$$

demonstrating that the number vectors  $\bar{\mathbf{x}}_i$  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  $\bar{\mathbf{x}}_i$  share a support  $\bar{S}$  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) [17] follows. Since the right-hand side of [17] is less than one, by [11] we have the contradiction  $k = \dim(\mathcal{A}_S) \leq \dim(\mathcal{B}_{\bar{S}}) \leq |\bar{S}|$ .

The total number of  $(k-1)$ -sparse vectors  $\bar{\mathbf{x}}_i$  thus can not exceed  $|\mathcal{H}|(k-1)\binom{\bar{m}}{k-1}$ . By [18], no more than  $|\mathcal{H}|k(k-1)\binom{\bar{m}}{k-1}$  vectors  $\bar{\mathbf{x}}_i$  are not  $k$ -sparse. Since for every  $S \in \mathcal{H}$  there are over  $(k-1)\left[\binom{\bar{m}}{k} + |\mathcal{H}|k\binom{\bar{m}}{k-1}\right]$  vectors  $\mathbf{x}_i$  supported there, it must be that more than  $(k-1)\binom{\bar{m}}{k}$  of them have corresponding  $\bar{\mathbf{x}}_i$  that are  $k$ -sparse. The result now follows from Thm. 1.  $\square$

*Proof (sketch) of Thm. 3.* Let  $\mathbf{M}$  be the matrix with columns  $\mathbf{A}\mathbf{x}_i$ ,  $i \in [N]$ . Consider the polynomial in the indeterminate entries of  $\mathbf{A}$  and  $\mathbf{x}_i$ :

$$g(\mathbf{A}, \{\mathbf{x}_i\}_{i=1}^N) := \prod_{S \in \binom{[N]}{2k}} \sum_{S' \in \binom{[n]}{2k}} (\det \mathbf{M}_{S', S})^2,$$

with notation as in Sec. 1. It can be checked that when  $g$  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 on  $\mathbf{A}$ ).  $\square$

*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  $\nu_\ell$  is absolutely continuous with respect to  $\mu$  for all  $\ell \in [p]$ , where  $\mu$  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.  $\square$

### 3. Discussion

The goal of this work was to explain the emergence of characteristic representations from sparse linear 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 hundreds (if not thousands) of publications on the topic by demonstrating very general, deterministic conditions under which the identification of parameters in this parsimonious model is not only possible but also robust to the inevitable uncertainty permeating measurement and model choice.

Specifically, 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 sparse codes consistent with the data are equivalent up to inherent relabeling/scaling ambiguities and a 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 perhaps the only published theory of neurally-plausible bottleneck communication in the brain: that sparse linear coding in a compressed space of neural activity can recover sparse codes sent through a randomly-constructed (but unknown) noisy wiring bottleneck (9).<sup>||</sup>

Beyond an original extension of existing noiseless guarantees (17) 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 new 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.

A technical difficulty in proving Thm. 1 is the absence of any assumption at all on dictionaries in solutions to Prob. 1. We sought such a 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 (17); indeed, even certifying a dictionary has this property is NP-hard (21).

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 (30), image compression (31), and, more recently, the design of an ultrafast camera (32). Given such effective uses of compressed sensing, it is only a matter of time before these systems incorporate dictionary learning to encode and process data (e.g., in a device that learns structure from motion (33)). 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

<sup>||</sup>We refer the reader to (29) for more on interactions between dictionary learning and neuroscience.

identical systems.\*\* 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 linear coding model, confirmation of which forms the core of our theoretical findings.

**ACKNOWLEDGMENTS.** We thank Friedrich Sommer and Darren Rhea for early thoughts, and Ian Morris for posting [12] online.

1. Y. Sato. A method of self-recovering equalization for multilevel amplitude-modulation systems. *IEEE Trans. Commun.*, 23(6):679–682, 1975.
2. Zheng Zhang, Yong Xu, Jian Yang, Xuelong Li, and David Zhang. A survey of sparse representation: algorithms and applications. *Access, IEEE*, 3:490–530, 2015.
3. Z. Wang, J. Yang, H. Zhang, Z. Wang, Y. Yang, D. Liu, and T. Huang. *Sparse coding and its applications in computer vision*. World Scientific, 2015.
4. B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.
5. J. Hurri, A. Hyvärinen, J. Karhunen, and E. Oja. Image feature extraction using independent component analysis. In *Proc. NORSIG '96 (Nordic Signal Proc. Symposium)*, pages 475–478, 1996.
6. A. Bell and T. Sejnowski. The “independent components” of natural scenes are edge filters. *Vision Res.*, 37(23):3327–3338, 1997.
7. J. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R Soc. Lond. [Biol.]*, 265(1394): 359–366, 1998.
8. D. Donoho and A. Flesia. Can recent innovations in harmonic analysis ‘explain’ key findings in natural image statistics? *Network Comp. Neural*, 12(3):371–393, 2001.
9. Guy Isely, Christopher Hillar, and Fritz Sommer. Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication. In *Adv. Neural Inf. Process. Syst.*, pages 910–918, 2010.
10. J. Hughes, D. Graham, and D. Rockmore. Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. *Proc. Natl. Acad. Sci.*, 107(4): 1279–1283, 2010.
11. B. Olshausen and M. DeWeese. Applied mathematics: The statistics of style. *Nature*, 463(7284):1027–1028, 2010.
12. J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Trans. Inf. Theory*, pages 853 – 884, 2016.
13. Jacques Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13(49-52):28, 1902.
14. Y. Li, A. Cichocki, and S.-I. Amari. Analysis of sparse representation and blind source separation. *Neural Comput.*, 16(6):1193–1234, 2004.
15. P. Georgiev, F. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Trans. Neural Netw.*, 16:992–996, 2005.
16. M. Aharon, M. Elad, and A. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra Appl.*, 416(1):48–67, 2006.
17. C. Hillar and F. Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *IEEE Trans. Inf. Theory*, 61(11):6290–6297, 2015.
18. J. Grcar. A matrix lower bound. *Linear Algebra Appl.*, 433(1):203–220, 2010.
19. Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1945–1959, 2005.
20. J. Blanchard, C. Cartis, and J. Tanner. Compressed sensing: How sharp is the restricted isometry property? *SIAM Rev.*, 53(1):105–125, 2011.
21. A. Tillmann and M. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inf. Theory*, 60(2):1248–1259, 2014.
22. M. Rehn and F. Sommer. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J. Comput. Neurosci.*, 22(2):135–146, 2007.
23. C. Rozell, D. Johnson, R. Baraniuk, and B. Olshausen. Neurally plausible sparse coding via thresholding and local competition. *Neural Comput.*, 20(10):2526–2563, 2008.
24. E. Arias-Castro, E. Candes, and M. Davenport. On the fundamental limits of adaptive sensing. *IEEE Trans. Inf. Theory*, 59(1):472–481, 2013.
25. A. Tillmann. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Process. Lett.*, 22(1):45–49, 2015.
26. Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.
27. I. Morris. A rapidly-converging lower bound for the joint spectral radius via multiplicative ergodic theory. *Adv. Math.*, 225(6):3425–3445, 2010.
28. F. Deutsch. *Best approximation in inner product spaces*. Springer Science & Business Media, 2012.
29. Surya Ganguli and Haim Sompolsky. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.*, 35:485–508, 2012.
30. M. Lustig, D. Donoho, J. Santos, and J. Pauly. Compressed sensing MRI. *IEEE Signal Process. Mag.*, 25(2):72–82, 2008.
31. M. Duarte, M. Davenport, D. Takbar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):83–91, March 2008.
32. L. Gao, J. Liang, C. Li, and L. Wang. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature*, 516(7529):74–77, 2014.

\*\*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 [34].

33. Chen Kong and Simon Lucey. Prior-less compressible structure from motion. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 4123–4131, 2016.
34. I Goodfellow, J Shlens, and C Szegedy. Explaining and harnessing adversarial examples. *Proc. International Conference on Learning Representations*, 11 pp., 2014.

## 4. Appendix

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

**Lemma 2.** *If  $f : U \rightarrow V$  is an injective function, then  $f(\cap_{i=1}^{\ell} U_i) = \cap_{i=1}^{\ell} f(U_i)$  for any  $U_1, \dots, U_{\ell} \subseteq U$ .*

*Proof.* By induction it is enough to prove the case  $|n| = 2$ , but this case follows directly from the assumption.  $\square$

**Lemma 3.** *Fix  $k \geq 2$ . Let  $V_1, \dots, V_k$  be subspaces of  $\mathbb{R}^m$  and set  $V = \cap_{i=1}^k V_i$ . For  $\mathbf{x} \in \mathbb{R}^m$ , we have (where  $\xi$  is given in Def. 4):*

$$\text{dist}(\mathbf{x}, V) \leq \frac{1}{\xi(\{V_i\}_{i=1}^k)} \sum_{i=1}^k \text{dist}(\mathbf{x}, V_i). \quad [19]$$

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

$$\begin{aligned} \|\mathbf{x} - \Pi_V \mathbf{x}\|_2 &\leq \|\mathbf{x} - \Pi_{V_k} \mathbf{x}\|_2 + \|\Pi_{V_k} \mathbf{x} - \Pi_{V_k} \Pi_{V_{k-1}} \mathbf{x}\|_2 \\ &\quad + \dots + \|\Pi_{V_k} \Pi_{V_{k-1}} \dots \Pi_{V_1} \mathbf{x} - \Pi_V \mathbf{x}\|_2 \\ &\leq \|\Pi_{V_k} \dots \Pi_{V_1} \mathbf{x} - \Pi_V \mathbf{x}\|_2 + \sum_{\ell=1}^k \|\mathbf{x} - \Pi_{V_{\ell}} \mathbf{x}\|_2, \quad [20] \end{aligned}$$

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

The desired result, [19], now follows by bounding the second term on the right-hand side using the following fact ([28, Thm. 9.33]):

$$\|\Pi_{V_k} \Pi_{V_{k-1}} \dots \Pi_{V_1} \mathbf{x} - \Pi_V \mathbf{x}\|_2 \leq z \|\mathbf{x}\|_2, \quad [21]$$

for  $z^2 = 1 - \prod_{\ell=1}^{k-1} (1 - z_{\ell}^2)$  and  $z_{\ell} = \cos \theta(V_{\ell}, \cap_{s=\ell+1}^k V_s)$ , recalling  $\theta$  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  $\|\Pi_{V_k} \dots \Pi_{V_1} \mathbf{x} - \Pi_V \mathbf{x}\|_2$  satisfies:

$$\|(\Pi_{V_k} \dots \Pi_{V_1} - \Pi_V)(\mathbf{x} - \Pi_V \mathbf{x})\|_2 \leq z \|\mathbf{x} - \Pi_V \mathbf{x}\|_2.$$

Finally, substituting this into [20] and rearranging produces [19] after replacing  $1 - z$  with  $\xi(\{V_i\}_{i=1}^k)$ .  $\square$

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

$$|\cap \pi(\mathcal{G})| \leq |\cap \mathcal{G}|, \quad \text{for } \mathcal{G} \in \binom{\mathcal{H}}{r} \cup \binom{\mathcal{H}}{r+1}, \quad [22]$$

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

*Proof.* Consider the collection of pairs:  $T_1 := \{(i, S) : i \in \pi(S), S \in \mathcal{H}\}$ , which number  $|T_1| = \sum_{S \in \mathcal{H}} |\pi(S)| \geq \sum_{S \in \mathcal{H}} |S| = \sum_{i \in [m]} \deg_{\mathcal{H}}(i) = mr$ . Note that [22] implies  $|T_1| \leq \bar{m}r$ , since otherwise pigeonholing the elements of  $T_1$  with respect to their possible first indices  $[\bar{m}]$  would imply  $\deg_{\mathcal{H}}(i) > r$  for some  $i \in [m]$ . Thus  $\bar{m} \geq m$ .

If  $\bar{m} < mr/(r-1)$  then  $p := \bar{m} - r(\bar{m} - m)$  is positive and  $|T_1| \geq \bar{m}(r-1) + p$ . Pigeonholing  $T_1$  into  $[\bar{m}]$  again, there must be  $r$  elements of  $T_1$  sharing the same first index. By [22], the intersection of the set  $\mathcal{G}_1$  of their second indices is nonempty; hence,  $|\cap \mathcal{G}_1| = 1$  by  $r$ -regularity and the SIP. Since  $\cap \pi(\mathcal{G}_1)$  is nonempty, applying [22] again gives  $\cap \pi(\mathcal{G}_1) = \{i_1\}$  for some  $i_1 \in [\bar{m}]$ . If  $p = 1$ , then we are done. Otherwise, define  $T_2 := T_1 \setminus \{(i, S) \in T_1 : i = i_1\}$ , which contains  $|T_2| = |T_1| - r \geq (\bar{m}-1)(r-1) + (p-1)$  ordered pairs having  $\bar{m}-1$  distinct first indices. Repeating the above arguments produces a (necessarily, by  $r$ -regularity and the SIP) distinct  $\mathcal{G}_2 \in \binom{\mathcal{H}}{r}$  and corresponding  $i_2 \in [\bar{m}]$ . Finally, iterating this  $p$  times in total yields the set of singletons  $J = \{\cap \mathcal{G}_1, \dots, \cap \mathcal{G}_p\} \subseteq [m]$ .  $\square$

*Proof of Lem. 1.* We begin by showing that  $\dim(\mathbf{B}_{\pi(S)}) = \dim(\mathbf{A}_S)$  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  $\|\mathbf{x}\|_1 \leq \sqrt{2}\|\mathbf{x}\|_2$  for all 2-sparse  $\mathbf{x}$ , by [5] we have  $L_2(\mathbf{A}) \leq \max_j \|\mathbf{A}_j\|_2$  and therefore (as  $\xi \leq 1$ ), the right-hand side of [13] is less than one. It follows from [11] that  $|\pi(S)| \geq \dim(\mathbf{B}_{\pi(S)}) \geq \dim(\mathbf{A}_S) = |S|$ , with the equality by injectivity of  $\mathbf{A}$ . Since  $|S| = |\pi(S)|$ , the required fact follows. Note that  $\mathbf{B}_{\pi(S)}$  therefore has full column-rank for all  $S \in \mathcal{H}$ .

We next show that [22] holds. Fix  $\mathcal{G} \subseteq \binom{\mathcal{H}}{r} \cup \binom{\mathcal{H}}{r+1}$ . Since  $\mathbf{B}_{\cap \pi(\mathcal{G})} \subseteq \cap_{S \in \mathcal{G}} \mathbf{B}_{\pi(S)}$ , if  $\cap_{S \in \mathcal{G}} \mathbf{B}_{\pi(S)} = \{\mathbf{0}\}$ , then we must have  $|\cap \pi(\mathcal{G})| = 0$  (as the columns of  $\mathbf{B}_{\pi(S)}$  are linearly independent) and [22] is trivially true. Suppose then that the intersection is not the zero vector. Keeping in mind that  $d(U', V) \leq d(U, V)$  whenever  $U' \subseteq U$  and applying (in order) Lem. 2, Lem. 3, [12], [13], and [16] gives:

$$\begin{aligned} d(\mathbf{B}_{\cap \pi(\mathcal{G})}, \mathbf{A}_{\cap \mathcal{G}}) &\leq d(\cap_{S \in \mathcal{G}} \mathbf{B}_{\pi(S)}, \cap_{S \in \mathcal{G}} \mathbf{A}_S) \\ &\leq \sum_{T \in \mathcal{G}} \frac{d(\cap_{S \in \mathcal{G}} \mathbf{B}_{\pi(S)}, \mathbf{A}_T)}{\xi(\{\mathbf{A}_S\}_{S \in \mathcal{G}})} \leq \sum_{T \in \mathcal{G}} \frac{d(\mathbf{B}_{\pi(T)}, \mathbf{A}_T)}{\xi(\{\mathbf{A}_S\}_{S \in \mathcal{G}})} \\ &\leq \frac{|\mathcal{G}| \varepsilon}{\xi(\{\mathbf{A}_S\}_{S \in \mathcal{G}})} \leq \frac{C_2 \varepsilon}{\max_j \|\mathbf{A}_j\|_2}. \end{aligned} \quad [23]$$

Since  $C_2 \varepsilon < L_2(\mathbf{A}) \leq \max_j \|\mathbf{A}_j\|_2$ , we have by [11] that  $\dim(\mathbf{B}_{\cap \pi(\mathcal{G})}) \leq \dim(\mathbf{A}_{\cap \mathcal{G}})$ , and [22] follows from the full column-rank of  $\mathbf{A}_S$  and  $\mathbf{B}_{\pi(S)}$  for all  $S \in \mathcal{G}$ .

Now, by Lem. 4, the association  $i \mapsto \cap \pi(\sigma(i))$  defines an injective map  $\bar{\pi} : J \rightarrow [\bar{m}]$  for some  $J \subseteq [m]$  of size  $\bar{m} - r(\bar{m} - m)$ , and we can be sure that  $\mathbf{B}_{\bar{\pi}(i)} \neq \mathbf{0}$  for all  $i \in J$  since the columns of  $\mathbf{B}_{\pi(S)}$  are linearly independent for all  $S \in \mathcal{H}$ . It then follows from [12] and [23] that  $d(\mathbf{A}_i, \mathbf{B}_{\bar{\pi}(i)}) \leq C_2 \varepsilon / \max_i \|\mathbf{A}_i\|_2$  for all  $i \in J$ . Fixing  $\bar{\varepsilon} = C_2 \varepsilon$  and letting  $c_i = \|\mathbf{A}_i\|_2^{-1}$ , we thus have that for every basis vector  $\mathbf{e}_i \in \mathbb{R}^m$  with  $i \in J$  there exists some  $\bar{c}_i \in \mathbb{R}$  such that  $\|c_i \mathbf{A} \mathbf{e}_i - \bar{c}_i \mathbf{B} \mathbf{e}_{\bar{\pi}(i)}\|_2 \leq \bar{\varepsilon} < L_2(\mathbf{A}) \min_{i \in J} |c_i|$ . But this is exactly the supposition in [9], and the result follows from the case  $k = 1$  in Sec. 2 applied to the submatrix  $\mathbf{A}_J$ .  $\square$