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

Charles J. Garfinkle[a,b, 1] **and Christopher J. Hillar**[a,1]

[a]Redwood Center for Theoretical Neuroscience, Berkeley, CA, USA; [b]Helen Wills Neuroscience Institute, UC Berkeley

Dictionary learning for sparse linear coding has exposed some characteristic properties of many natural signals. However, universal theorems which guarantee the consistency of estimation in this model are 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, compressive sensing, inverse problems

Blind source separation is a classical problem in signal processing (1). A common modern assumption is that 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 $k < m \ll N$ (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 natural image training set.

Motivated by these discoveries and earlier work in the theory of neural communication (9, 10), 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 (11, 12), 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 (13, Sec. I-E) for a summary of the state-of-the-art). Few theorems, however, can be cited to explain this uniqueness phenomenon more generally.

Here, we prove very generally that uniqueness and stability in sparse linear coding is an expected property of the model. More specifically, dictionaries that preserve sparse codes (i.e., 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 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 (14)) given enough data (Cor. 3). Moreover, these guarantees hold without assuming the recovered matrix satisfies a spark condition, even when the number m of dictionary elements is unknown. The explicit, algorithm-independent criteria we provide should be a useful theoretical tool.

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

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

for *k-sparse* $\mathbf{x}_i \in \mathbb{R}^m$ having at most k 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 precise mathematical problem addressed here is the following.

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

Note that any particular solution $(\mathbf{B}, \bar{\mathbf{x}}_1 \ldots, \bar{\mathbf{x}}_N)$ to this problem gives rise to an orbit of equivalent solutions $(\mathbf{BPD}, \mathbf{D}^{-1}\mathbf{P}^\top\bar{\mathbf{x}}_1, \ldots, \mathbf{D}^{-1}\mathbf{P}^\top\bar{\mathbf{x}}_N)$, where $\mathbf{P}$ is any permutation matrix and $\mathbf{D}$ any invertible diagonal. Previous theoretical work addressing the noiseless case $\eta = 0$ (e.g., (15–18)) 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 compressive sensing:

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

which would be, in any case, necessary for uniqueness. Our concern here is solution stability with respect to noise.

Definition 1. *Fix* $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \subset \mathbb{R}^n$*. We say* Y *has a* k***-sparse representation in*** $\mathbb{R}^m$ *if there exists a matrix* $\mathbf{A}$ *and* k*-sparse* $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ *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* $\varepsilon(\delta_1, \delta_2) \geq 0$ *(with* $\varepsilon > 0$ *when* $\delta_1, \delta_2 > 0$*) such that if* $\mathbf{B}$ *and* k*-sparse* $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_N \in \mathbb{R}^m$ *satisfy:*

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

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

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

Significance Statement

Many naturally occurring 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 may help explain the observed universality of emergent representations in sparse models of some natural phenomena.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: chaz@berkeley.edu, chillar@msri.org

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 δ_1, δ_2 to be the desired recovery accuracy in Eq. (3). Consider now any dataset Z generated as in Eq. (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 $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_N \in \mathbb{R}^m$ solving Prob. 1 are necessarily close to the original dictionary $\mathbf{A}$ and sparse codes $\mathbf{x}_i$.

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). We also give a simple argument extending our guarantees to the following more common optimization formulation of the dictionary learning problem (Cor. 2).

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

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

All other proofs are relegated to the Section 4. Finally, we present several applications in Discussion Sec. 3.

1. Results

Before stating our main results, we first identify combinatorial criteria on the support sets of generating codes that imply stable sparse representations. Letting $\{1, \ldots, m\}$ 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 $E \subseteq 2^{[m]}$ on vertices $[m]$ is *k-uniform* when in fact $E \subseteq \binom{[m]}{k}$. The *degree* $\deg(i)$ of a node $i \in [m]$ is the number of elements in E which contain i, and we say E is *regular* when for some r we have $\deg(i) = r$ for all i (for given r, we say E is *r-regular*).

Definition 2. *Given $E \subseteq 2^{[m]}$, the* ***star*** *$H(i)$ at i is the set of $S \in E$ with $i \in S$. We say E has the* ***singleton intersection property (SIP)*** *when $\cap H(i) = \{i\}$ for all $i \in [m]$.*

Next, we describe a quantitative version of the spark condition. The *lower bound* L of a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ is the largest number α such that $\|\mathbf{M}\mathbf{x}\|_2 \geq \alpha\|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{R}^m$ (19). By compactness of the unit sphere, every injective linear map has a nonzero lower bound; hence, if $\mathbf{M}$ satisfies Eq. (2), then each submatrix formed from $2k$ of its columns or less has a nonzero lower bound. This motivates the following definition:

$$L_E(\mathbf{M}) := \inf \left\{ \frac{\|\mathbf{M}_{S \cup S'}\mathbf{x}\|_2}{\sqrt{2k}\|\mathbf{x}\|_2} : \mathbf{x} \in \mathbb{R}^{|S \cup S'|}, \quad S, S' \in E \right\}$$

and we write $L_{2k} := L_E$ when $E = \binom{[m]}{k}$. Note that $L = L_{\{[m]\}}$ whereas $L_{[m]} = L_2$. [TODO: Note when $L_2 > 0$ is implied by $L_E > 0$.]

Clearly, for any $\mathbf{M}$ satisfying Eq. (2), we have $L_{k'}(\mathbf{M}) > 0$ for $k' \leq 2k$. We note that the quantity $1 - \sqrt{k}L_k(\mathbf{M})$ is also known in the compressive sensing literature as the (asymmetric) lower restricted isometry constant (20).

A vector $\mathbf{x}$ is said to be *supported* on $S \subseteq [m]$ when $\mathbf{x} \in \mathrm{Span}(\{\mathbf{e}_j\}_{j \in S})$, where $\mathbf{e}_j$ are the standard basis in $\mathbb{R}^m$. For any index set J, denote by by $\mathbf{x}^J$ the subvector formed from the entries of $\mathbf{x}$ indexed by J, and similarly by $\mathbf{M}_J$ the submatrix formed by the columns of $\mathbf{M}$ indexed by J, where we set $\mathrm{Span}\{\mathbf{M}_\emptyset\} := \{\mathbf{0}\}$. 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.

Theorem 1. *Fix a real $n \times m$ matrix $\mathbf{A}$ with $L_E(\mathbf{A}) > 0$ for some r-regular $E \subseteq \binom{[m]}{k}$ with the SIP. If the sequence $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ contains, for each $S \in E$, more than $(k-1)\binom{m}{k}$ k-sparse vectors in general linear position supported on S, then there is a constant* $C_1 > 0$ for which the following holds for all $\varepsilon < L_2(\mathbf{A})/C_1$:*

Every real $n \times \bar{m}$ matrix $\mathbf{B}$ for which there exist k-sparse $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_N$ 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 $\bar{m} < mr/(r-1)$,

$$\|(\mathbf{A}_J)_j - \mathbf{B}_{\bar{J}}\mathbf{P}\mathbf{D}_j\|_2 \leq C_1\varepsilon \quad \textit{for all } j \in J, \qquad [5]$$

for some non-empty $J, \bar{J}$ of size $\bar{m} - r(\bar{m} - m)$, permutation $\mathbf{P}$ and invertible diagonal $\mathbf{D}$.

Moreover, if in fact $L_{2k}(\mathbf{A}) > 0$ 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:

$$\|\mathbf{x}_i^J - \mathbf{D}^{-1}\mathbf{P}^\top \bar{\mathbf{x}}_i^{\bar{J}}\|_1 \leq \left(\frac{1 + C_1\|\mathbf{x}_i^J\|_1}{L_{2k}(\mathbf{A}) - C_1\varepsilon} \right) \varepsilon, \quad \textit{for } i \in [N]. \qquad [6]$$

To be clear, Thm. 1 says that the smaller the difference $\bar{m} - m$, the more columns and coefficients of the original dictionary $\mathbf{A}$ and codes $\mathbf{x}_i$ contained (up to noise) in the appropriately scaled dictionary $\mathbf{B}$ and codes $\bar{\mathbf{x}}_i$. The implication in the case $\bar{m} = m$ is that $Y = \{\mathbf{A}\mathbf{x}_1, \ldots, \mathbf{A}\mathbf{x}_N\}$ has a stable k-sparse representation in $\mathbb{R}^m$, with Eq. (3) guaranteed provided ε in Def. 1 does not exceed:

$$\varepsilon(\delta_1, \delta_2) := \min \left\{ \frac{\delta_1}{C_1}, \frac{\delta_2 L_{2k}(\mathbf{A})}{1 + C_1\left(\max_{i \in [N]} \|\mathbf{x}_i\|_1 + \delta_2\right)} \right\}. \qquad [7]$$

Regarding the assumptions of Thm. 1, it is easy to verify that for every $k < m$, there is a regular k-uniform hypergraph that satisfies the SIP; for instance, take the consecutive intervals of length k in some cyclic order on $[m]$. In fact, even if E is not regular or only partially satisfies the SIP, a relation between $\bar{m}$ and the degree sequence of nodes in E may give the indices $J \subseteq [m]$. For sake of brevity, we delay to the next section a clear statement of this more general result.

It also happens that producing sparse codes $\mathbf{x}_i$ in general linear position is straightforward with a "Vandermonde" matrix construction (e.g., see (18)). We therefore have:

Corollary 1. *Given n, m, $k < m$, and a regular hypergraph $E \subseteq \binom{[m]}{k}$ with the SIP, there are $N = |E|\left[(k-1)\binom{m}{k} + 1\right]$ vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ such that every matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ satisfying* Eq. (2) *generates a dataset $Y = \{\mathbf{A}\mathbf{x}_1, \ldots, \mathbf{A}\mathbf{x}_N\}$ with a stable k-sparse representation in $\mathbb{R}^m$.*

As already mentioned, there exist k-uniform regular hypergraphs E with the SIP having cardinality $|E| = m$, implying the lower bound for sample size N from the introduction. In many cases, however, the SIP can be achieved with far fewer supports; for example, when $k = \sqrt{m}$, take E to be the $2k$ rows and columns formed by arranging $[m]$ in a square grid.

There are other less direct consequences of Thm. 1. For instance, it has the following implications for the optimization problem posed in Prob. 2:

*We delay defining the constant C_1 until Section 2 (Eq. (16)).

Corollary 2. *If all of the assumptions of Thm. 1 hold with more than* $(k-1)\left[\binom{\bar{m}}{k}+|E|k\binom{\bar{m}}{k-1}\right]$ *vectors* $\mathbf{x}_i$ *supported on each* $S\in E$*, then all solutions to Prob. 2 necessarily satisfy the implications* Eq. (5) *and* Eq. (6) *of Thm. 1.*

Another extension of Thm. 1 follows 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 Eq. (2) if and only if the following polynomial is nonzero:

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

where for any $S'\in\binom{[n]}{k}$ and $S\in\binom{[m]}{k}$, 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 an analytic function, if there exists one substitution of a real matrix $\mathbf{A}$ such that $f(\mathbf{A})\neq 0$ then the zeroes of f in fact form a set of 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 $[\gamma_i^j]_{i,j=1}^{k,m}$ with distinct γ_i (so that each term in the product above is nonzero). Hence, almost every real $n\times m$ matrix with $n\geq\min(2k,m)$ satisfies Eq. (2).

A similar phenomenon applies to datasets of vectors with a stable sparse representation. As in (18, 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 2. *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 on each* $S\in E$ *for some regular* $E\subseteq\binom{[m]}{k}$ *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 3. *Fix* $k<m$ *and* $n\geq\min(2k,m)$ *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* μ*. If more than* $(k-1)\binom{m}{k}$ *of the vectors* $\mathbf{x}_i$ *are supported on each* $S\in E$ *for a regular* $E\subseteq\binom{[m]}{k}$ *with the SIP, then* Y *has a stable* k*-sparse representation in* $\mathbb{R}^m$ *with probability one.*

Thus, choosing the dictionary and sparse codes "randomly" almost certainly generates data with stable sparse representations.

2. Proofs of Theorem 1 and Corollary 2

As mentioned in the previous section, Thm. 1 is a particular case of a more general result that forgoes the assumption that $E\subseteq\binom{[m]}{k}$ is regular and satisfies the SIP. If instead we require only that the stars $\cap H(i)$ intersect at singletons for all $i\leq q$ (assuming that the nodes of E are labeled in some order of non-increasing degree), we have that $\bar{m}\geq k|E|/\deg(1)$ and, provided $\bar{m}<k|E|/(\deg(1)-1)$, the non-empty submatrix J is of size equal to the largest number p satisfying:

$$\sum_{i=\ell}^{m}\deg(i)>(\bar{m}+1-\ell)(\deg(\ell)-1)\quad\text{for all }\ell\leq p\leq q. \tag{8}$$

Specifically, J consists of the union of the set of all nodes of degree exceeding $\deg(p)$ and some subset of those nodes with degrees equal to $\deg(p)$. For the benefit of the reader, we do not prove this more general result below; it can be gleaned by examining how exactly Lemma 4 is incorporated into the proof of Lem. 1.

We now begin our proof of Thm. 1 by showing how dictionary recovery Eq. (5) already implies sparse code recovery Eq. (6) when $L_{2k}(\mathbf{A})>0$ and $\varepsilon<L_{2k}(\mathbf{A})/C_1$, temporarily assuming (without loss of generality) that $\bar{m}=m$. First, note that $\|\mathbf{x}\|_1\leq\sqrt{k}\|\mathbf{x}\|_2$ for k-sparse $\mathbf{x}\in\mathbb{R}^m$, which by definition of L_{2k} implies the following frequently applied inequality:

$$L_{2k}(\mathbf{A})\leq\frac{\|\mathbf{A}\mathbf{x}\|_2}{\sqrt{2k}\|\mathbf{x}\|_2}\leq\max_{j\in[m]}\|\mathbf{A}_j\|_2. \tag{9}$$

For k-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,

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

Hence, $L_{2k}(\mathbf{BPD})\geq L_{2k}(\mathbf{A})-C_1\varepsilon>0$. Eq. (6) 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 Eq. (5), which we now finally establish, first in the important special case $k=1$.

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 may take any $C_1\geq 1/\min_{\ell\in[m]}|c_\ell|$.

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

$$\|c_j\mathbf{A}_j-\bar{c}_j\mathbf{B}_{\pi(j)}\|_2\leq\varepsilon,\quad\text{for }j\in[m]. \tag{10}$$

Note that if $\bar{c}_j=0$, then $\|c_j\mathbf{A}_j\|_2\leq\varepsilon$ implies from Eq. (9) that $|c_j|<\min_{\ell\in[m]}|c_\ell|$, a contradiction. Thus, $\bar{c}_j\neq 0$, $j\in[m]$.

We now show that π is injective (in particular, a permutation if $\bar{m}=m$). Suppose that $\pi(i)=\pi(j)=\ell$ for some $i\neq j$ and ℓ. 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:

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

which contradicts the bound $\varepsilon<L_2(\mathbf{A})/C_1$. Hence, the map π is injective and therefore $\bar{m}\geq m$. Setting $\bar{J}=\pi([m])$ and letting $P=\left(\mathbf{e}_{\pi(1)}\cdots\mathbf{e}_{\pi(m)}\right)$ and $D=\operatorname{diag}(\frac{\bar{c}_1}{c_1},\dots,\frac{\bar{c}_m}{c_m})$, we see that Eq. (10) becomes, for all $j\in[m]$:

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

□

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

$$d(U, V) := \sup_{\mathbf{u} \in U,\ \|\mathbf{u}\|_2 \leq 1} \operatorname{dist}(\mathbf{u}, V). \qquad [11]$$

We note the following facts about d. For subspaces $U \subseteq U', V \subseteq \mathbb{R}^m$, we have $d(U, V) \leq d(U', V)$ and (22, Cor. 2.6):

$$d(U, V) < 1 \implies \dim(U) \leq \dim(V). \qquad [12]$$

Also, from (23, Lem. 3.2), we have:

$$\dim(U) = \dim(V) \implies d(U, V) = d(V, U). \qquad [13]$$

Our result in combinatorial matrix analysis is the following.

Lemma 1. *Suppose $\mathbf{A} \in \mathbb{R}^{n \times m}$ has $L_E(\mathbf{A}) > 0$ for some r-regular $E \subseteq \binom{[m]}{k}$ 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 $\mathbf{B} \in \mathbb{R}^{n \times \bar{m}}$ and map $\pi : E \mapsto \binom{[\bar{m}]}{k}$ we have:

$$d(\operatorname{Span}\{\mathbf{A}_S\}, \operatorname{Span}\{\mathbf{B}_{\pi(S)}\}) \leq \varepsilon, \quad \text{for } S \in E, \qquad [14]$$

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

$$\|(\mathbf{A}_J)_j - \mathbf{B}_{\bar{J}} \mathbf{P} \mathbf{D}_j\|_2 \leq C_2 \varepsilon, \quad \text{for } j \in J, \qquad [15]$$

for some non-empty $J, \bar{J}$ of size $\bar{m} - r(\bar{m} - m)$.

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

$$C_1(\mathbf{A}, \{\mathbf{x}_i\}_{i=1}^N, E) := \frac{C_2(\mathbf{A}, E)}{\min_{S \in E} L_k(\mathbf{A}\mathbf{X}_{I(S)})}. \qquad [16]$$

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 the support of $\mathbf{x}_i$ is contained in S.

The constant $C_2 = C_2(\mathbf{A}, E)$ is given in turn below, in terms of one used in (24) to analyze the convergence of the alternating projections algorithm for projecting a point onto an intersection of subspaces. We use it 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 $V_1, \ldots, V_\ell \subseteq \mathbb{R}^m$, set $r := 1$ when $\ell = 1$ and define for $\ell \geq 2$:*

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

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 (24, Def. 9.4):

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

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

[‡] We modify the quantity in (24) in this way since the subspace ordering is irrelevant to our purpose.

Note that $\theta \in (0, \frac{\pi}{2}]$ implies $0 < r \leq 1$.[§] The constant C_2 in Lem. 1 can then be expressed as:

$$C_2(\mathbf{A}, E) := \frac{2^{|E|} \max_{j \in [m]} \|\mathbf{A}_j\|_2}{\min_{F \subseteq E} r(\{\operatorname{Span}\{\mathbf{A}_S\}\}_{S \in F})}, \qquad [17]$$

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.

Proof of Thm. 1 for $k < m$. We shall show that for every $S \in E$ there is some $\bar{S} \in \binom{[\bar{m}]}{k}$ for which the distance $d(\operatorname{Span}\{\mathbf{A}_S\}, \operatorname{Span}\{\mathbf{B}_{\bar{S}}\})$ is controlled by ε. Applying Lem. 1 with the map π defined by $S \mapsto \bar{S}$ then completes the proof.

Since there are more than $(k-1)\binom{\bar{m}}{k}$ vectors $\mathbf{x}_i$ supported on S, by the pigeonhole principle there must be some $\bar{S} \in \binom{[\bar{m}]}{k}$ and set of k indices $K \subseteq I(S)$ such that the supports of all $\bar{\mathbf{x}}_i$ with $i \in K$ are subsets of $\bar{S}$.

It follows from the general linear position of the $\mathbf{x}_i$ and the linear independence of the columns of $\mathbf{A}_S$ 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 $\operatorname{Span}\{\mathbf{A}_S\}$. Fixing $\mathbf{0} \neq \mathbf{y} \in \operatorname{Span}\{\mathbf{A}_S\}$, 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 $\bar{\mathbf{y}} = \mathbf{B}\bar{\mathbf{X}}_K\mathbf{c} \in \operatorname{Span}\{\mathbf{B}_{\bar{S}}\}$, we have:

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

where the last inequality follows directly from the definition of L. From Def. 3 we have:

$$d(\operatorname{Span}\{\mathbf{A}_S\}, \operatorname{Span}\{\mathbf{B}_{\bar{S}}\}) \leq \frac{\varepsilon}{L(\mathbf{A}\mathbf{X}_K)} \leq \varepsilon \frac{C_1}{C_2}, \qquad [18]$$

where the second inequality follows from $L(\mathbf{A}\mathbf{X}_K) \geq L_k(\mathbf{A}\mathbf{X}_{I(S)})$ and Eq. (16). Since $\varepsilon < L_2(\mathbf{A})/C_1$, the result follows by Lem. 1, . □

Proof of Cor. 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 Eq. (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,

$$\sum_{p=k+1}^{\bar{m}} n_p \leq \sum_{p=k+1}^{\bar{m}} (p-k) n_p \leq \sum_{p=0}^{k} (k-p) n_p \leq k \sum_{p=0}^{k-1} n_p. \qquad [19]$$

We now show that no more than $(k-1)|E|$ of the $\bar{\mathbf{x}}_i$ share a support of size less than k. From previous arguments, for every $S \in E$, the columns of each $n \times k$ submatrix of $\mathbf{A}\mathbf{X}_{I(S)}$ form a basis for $\operatorname{Span}\{\mathbf{A}_S\}$. Suppose that more than $(k-1)|E|$ vectors $\bar{\mathbf{x}}_i$ share a support $\bar{S}$ of size $|\bar{S}| < k$. By the pigeonhole principle, there is some $S \in E$ supporting k or more of the corresponding $\mathbf{x}_i$; let them be indexed by $K \subseteq I(S)$. By the same argument as in the proof of Thm.1, we also have Eq. (18). Our bound on ε implies the right-hand side of Eq. (18) is less than one; hence, by Eq. (12) we have the contradiction:

$$k = \dim(\operatorname{Span}(\mathbf{A}_S)) \leq \dim(\operatorname{Span}\{\mathbf{B}_{\bar{S}}\}) \leq |\bar{S}|.$$

The total number of $(k-1)$-sparse vectors $\bar{\mathbf{x}}_i$ can thus no greater than $|E|(k-1)\binom{\bar{m}}{k-1}$. Taking Eq. (19) into account,

[§] We acknowledge the counter-intuitive property that $\theta = \pi/2$ when $U = V$ or $U \perp V$.

no more than $|E|k(k-1)\binom{\bar{m}}{k-1}$ vectors $\bar{\mathbf{x}}_i$ are *not* k-sparse. Since for every $S \in E$ there are over $(k-1)\left[\binom{\bar{m}}{k} + |E|k\binom{\bar{m}}{k-1}\right]$ vectors $\mathbf{x}_i$ supported there, it follows that more than $(k-1)\binom{\bar{m}}{k}$ of them must have corresponding $\bar{\mathbf{x}}_i$ that are also k-sparse. The result now follows directly from Thm. 1. □

3. Discussion

In this note, we generalized the approach of (18) to prove the stability of unique solutions to Probs. 1 and 2 while significantly reducing the known sample complexity. Our results justify the application of the sparse linear coding model to blind source separation problems, wherein the goal is to infer the generating dictionary and sparse codes from noisy measurements, Eq. (1). We also collect a set of useful mathematical tools and basic facts for future research. The main motivation for this work, however, was to understand how seemingly universal representations emerge from sparse coding models fit to natural data by a variety of methodologies. We elaborate on these applications below after discussing theoretical aspects.

What we have shown here is that the sparse linear coding model generally produces a *well-posed* inverse problem to be approximately solved by a numerical algorithm. This early concept of Hadamard (14) can be paraphrased as the idea that inferences from observations should be robust to the inevitable uncertainty in measurement. In other words, a small perturbation of the data should result in only slightly different inferred parameters. In this regard, for the sparse linear coding model we demonstrate a linear relationship, Eq. (7), between measurement noise and the deviation of any solution from the true parameters, with explicit constants expressed in terms of these parameters. Moreover, we show that even if the meta-parameter for the number of dictionary elements is overestimated, a subset of parameters may still be identifiable up to noise. It would therefore be of practical utility to determine the best possible dependence of ε on δ_1, δ_2 in Def.1 as well as the minimal requirements on the number and diversity of generating codes, and we hope that other researchers continue to improve and extend our results. We remark that our constants have been derived for deterministic "worst-case" noise, whereas the "effective" noise might be smaller when sampled from a distribution; in such cases, the constants will improve as well.

One notable component of our contribution is a combinatorial criteria (regular hypergraphs satisfying the singleton intersection property, Def. 2) for the support sets of sparse codes key to the identification of the dictionary. Fully understanding those combinatorial designs allowing for stable sparse representations is an interesting research area for the future. For instance, whether there is a support set size N that is polynomial in m, k gauranteeing the conclusions of Thm. 1 has implications for the computational complexity of Probs. 1 and 2 and related questions (25).

A technical difficulty in proving Thm. 1 was the absence of a spark condition assumption on solutions to Prob. 1. Although mathematically interesting that no such requirement is necessary, there are other reasons to seek out such a theoretical guarantee. For instance, it is difficult to ensure that an algorithm maintain a dictionary satisfying Eq. (2) at each iteration; indeed, even certifying a dictionary has this property is likely intractable given its NP-hardness (21).

In fact, uniqueness guarantees with minimal assumptions apply to all areas of data science and engineering that utilize learned sparse representations. For example, several groups have applied compressive sensing to signal processing tasks: MRI analysis (26), image compression (27), and, more recently, the design of an ultrafast camera (28). Given such effective uses of compressive 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 (29)). In these cases, assurances such as those offered by our theorems certify that different devices (with different initialization, samples, etc.) will learn equivalent representations given enough data from statistically identical systems.

In the field of theoretical neuroscience in particular, dictionary learning for sparse coding and related methods have recovered characteristic components of natural images (4–7) and sounds (30–32) that reproduce response properties of cortical neurons. Our results suggest that this correspondence could be due to the "universality" of sparse representations in natural data, an early mathematical idea in neural theory (33). Furthermore, they justify the hypothesis of (9, 10) that sparse codes passed through information bottlenecks in the brain are recovered from random projections via (unsupervised) biologically plausible sparse coding (e.g., (34–36)).

ACKNOWLEDGMENTS. We thank Friedrich Sommer and Darren Rhea for early thoughts, and Ian Morris for posting Eq. (13) online.

1. Sato Y (1975) A method of self-recovering equalization for multilevel amplitude-modulation systems. *IEEE Transactions on communications* 23(6):679–682.
2. Zhang Z, Xu Y, Yang J, Li X, Zhang D (2015) A survey of sparse representation: algorithms and applications. *Access, IEEE* 3:490–530.
3. Wang Z, et al. (2015) *Sparse Coding and its Applications in Computer Vision.* (World Scientific).
4. Olshausen B, Field D (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609.
5. Hurri J, Hyvärinen A, Karhunen J, Oja E (1996) Image feature extraction using independent component analysis in *Proc. NORSIG '96 (Nordic Signal Processing Symposium).* pp. 475–478.
6. Bell A, Sejnowski T (1997) The "independent components" of natural scenes are edge filters. *Vision Research* 37(23):3327–3338.
7. van Hateren J, van der Schaaf A (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265(1394):359–366.
8. Donoho D, Flesia A (2001) Can recent innovations in harmonic analysis 'explain' key findings in natural image statistics? *Network: computation in neural systems* 12(3):371–393.
9. Coulter W, Hillar C, Isley G, Sommer F (2010) Adaptive compressed sensing – a new class of self-organizing coding models for neuroscience in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* (IEEE), pp. 5494–5497.
10. Isely G, Hillar C, Sommer F (2010) Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication in *Advances in Neural Information Processing Systems.* pp. 910–918.
11. Hughes J, Graham D, Rockmore D (2010) Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. *Proc. of the National Academy of Sciences* 107(4):1279–1283.
12. Olshausen B, DeWeese M (2010) Applied mathematics: The statistics of style. *Nature* 463(7284):1027–1028.
13. Sun J, Qu Q, Wright J (2016) Complete dictionary recovery over the sphere I: Overview and the geometric picture. *Information Theory, IEEE Transactions on* pp. 853 – 884.
14. Hadamard J (1902) Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin* 13(49-52):28.
15. Li Y, Cichocki A, Amari SI (2004) Analysis of sparse representation and blind source separation. *Neural Computation* 16(6):1193–1234.
16. Georgiev P, Theis F, Cichocki A (2005) Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks* 16:992–996.
17. Aharon M, Elad M, Bruckstein A (2006) On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra and its Applications* 416(1):48–67.
18. Hillar C, Sommer F (2015) When can dictionary learning uniquely recover sparse data from subsamples? *Information Theory, IEEE Transactions on* 61(11):6290–6297.
19. Grcar J (2010) A matrix lower bound. *Linear Algebra and its Applications* 433(1):203–220.
20. Blanchard J, Cartis C, Tanner J (2011) Compressed sensing: How sharp is the restricted isometry property? *SIAM review* 53(1):105–125.
21. Tillmann A, Pfetsch M (2014) The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory* 60(2):1248–1259.

22. Kato T (2013) *Perturbation theory for linear operators.* (Springer Science & Business Media) Vol. 132.
23. Morris I (2010) A rapidly-converging lower bound for the joint spectral radius via multiplicative ergodic theory. *Advances in Mathematics* 225(6):3425–3445.
24. Deutsch F (2012) *Best approximation in inner product spaces.* (Springer Science & Business Media).
25. Tillmann A (2015) On the computational intractability of exact and approximate dictionary learning. *Signal Processing Letters, IEEE* 22(1):45–49.
26. Lustig M, Donoho D, Santos J, Pauly J (2008) Compressed sensing MRI. *IEEE Signal Processing Magazine* 25(2):72–82.
27. Duarte M, et al. (2008) Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine* 25(2):83–91.
28. Gao L, Liang J, Li C, Wang L (2014) Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature* 516(7529):74–77.
29. Kong C, Lucey S (2016) Prior-less compressible structure from motion in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* pp. 4123–4131.
30. Bell A, Sejnowski T (1996) Learning the higher-order structure of a natural sound. *Network: Computation in Neural Systems* 7(2):261–266.
31. Smith E, Lewicki M (2006) Efficient auditory coding. *Nature* 439(7079):978–982.
32. Carlson N, Ming V, DeWeese M (2012) Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput Biol* 8(7):e1002594.
33. Pitts W, McCulloch W (1947) How we know universals: the perception of auditory and visual forms. *The Bulletin of mathematical biophysics* 9(3):127–147.
34. Rehn M, Sommer F (2007) 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.
35. Rozell C, Johnson D, Baraniuk R, Olshausen B (2007) Neurally plausible sparse coding via thresholding and local competition. *Neural Computation.*
36. Hu T, Pehlevan C, Chklovskii D (2014) A Hebbian/anti-Hebbian network for online sparse dictionary learning derived from symmetric matrix factorization in *2014 48th Asilomar Conference on Signals, Systems and Computers.* (IEEE), pp. 613–619.
37. Folland G (2013) *Real analysis: modern techniques and their applications.* (John Wiley & Sons).

4. Appendix

We prove Lem. 1 after stating auxiliary lemmas and then sketch the proofs of Thm. 2 and Cor. 3.

Lemma 2. *If $f : U \to V$ is an injective function then:*

$$f\left(\cap_{i=1}^{\ell} U_i\right) = \cap_{i=1}^{\ell} f\left(U_i\right) \quad \text{for any } U_1, \ldots, U_\ell \subseteq U.$$

Proof. By induction it is enough to prove the case $|n| = 2$, but this case follows directly from the assumtion. □

Lemma 3. *Fix $k \geq 2$. Let $V_1, \ldots, V_k$ 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 r is given in Def. 4):*

$$\operatorname{dist}(\mathbf{x}, V) \leq \frac{1}{r(\{V_i\}_{i=1}^k)} \sum_{i=1}^{k} \operatorname{dist}(\mathbf{x}, V_i). \quad [20]$$

Proof. Recall the orthogonal projection onto a 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:

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

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, Eq. (20), now follows by bounding the second term on the right-hand side using the following fact (24, Thm. 9.33):

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

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 yields:

$$\begin{aligned}\|\Pi_{V_k}\cdots\Pi_{V_1}\mathbf{x} - \Pi_V\mathbf{x}\|_2 &= \|\left(\Pi_{V_k}\cdots\Pi_{V_1} - \Pi_V\right)(\mathbf{x} - \Pi_V\mathbf{x})\|_2 \\ &\leq z\|\mathbf{x} - \Pi_V\mathbf{x}\|_2.\end{aligned}$$

Finally, substituting this into Eq. (21) and rearranging produces Eq. (20) after replacing $1 - z$ with $r(\{V_i\}_{i=1}^k)$. □

Lemma 4. *Fix a hypergraph $E \subseteq 2^{[m]}$ with nodes labeled in order of non-increasing degree and for which $|\cap H(i)| = 1$ for all $i \leq q$. Fix $\bar{m}$ and let $p \leq q$ be the largest number satisfying:*

$$\sum_{i=\ell}^{m} \deg(i) > (\bar{m} + 1 - \ell)(\deg(\ell) - 1) \quad \text{for all } \ell \leq p. \quad [23]$$

If the map $\pi : E \to 2^{[\bar{m}]}$ has $\sum_{S\in E}(|\pi(S)| - |S|) \geq 0$ and:

$$|\cap\pi(F)| \leq |\cap F| \text{ for all } F \subseteq E, \quad [24]$$

then $\bar{m} \geq \sum_i \deg(i)/\deg(1)$, and if $\bar{m} < \sum_i \deg(i)/(\deg(1) - 1)$ then the association $i \mapsto \cap\pi(F(i))$ defines an injective map to $[\bar{m}]$ from some non-empty $J \subseteq [m]$ of size p consisting of the union of the set of all nodes of degree exceeding $\deg(p)$ and some set of nodes all having degree equal to $\deg(p)$. In particular, if E is r-regular then $|J| = \bar{m} - r(\bar{m} - m)$.

Proof. Consider the collection of pairs: $T_1 := \{(i, S) : i \in \pi(S), S \in E\}$, which number $|T_1| = \sum_{S\in E}|\pi(S)| \geq \sum_{S\in E}|S| = \sum_{i\in[m]}\deg(i)$. Note that assumption Eq. (24) implies $\bar{m} \geq |T_1|/\deg(1)$, since otherwise pigeonholing the elements of T_1 with respect to their set of possible first indices $[\bar{m}]$ would lead us to conclude that there are more than $\deg(1)$ sets in E sharing a common element.

By Eq. (23) and the upper bound on $\bar{m}$ we have $|T_1| > \bar{m}(\deg(1) - 1)$, which implies, again by the pigeonhole principle, that there must be at least $\deg(1)$ elements of T_1 sharing the same first index. By Eq. (24), the intersection of the set F_1 consisting of their second indices is non-empty. As $p \leq q$ and $\deg(1) \geq \deg(i)$ for all i, it must be that $\cap F_1| = 1$. Since $\cap\pi(F_1)$ is non-empty, applying Eq. (24) again implies $\cap\pi(F_1) = \{i_1\}$ for some $i_1 \in [m]$. If $p = 1$ then we are done. Otherwise, define $T_2 := T_1 \setminus \{(i, S) \in T_1 : i = i_1\}$, which contains $|T_2| = |T_1| - \deg(1) = \sum_{i=2}^{m}\deg(i)$ ordered pairs having $\bar{m} - 1$ distinct first indices. By Eq. (23) we have $|T_2| > (\bar{m} - 1)(\deg(2) - 1)$ and reiterating the above arguments produces a (necessarily) distinct index i_2. Iterating the arguments p times yields the set of singletons $J = \{\cap F_1, \ldots, \cap F_p\} \subseteq [m]$. □

Proof of Lem. 1. We begin by showing that:

$$d(\operatorname{Span}\{\mathbf{B}_{\pi(S)}\}, \operatorname{Span}\{\mathbf{A}_S\}) = d(\operatorname{Span}\{\mathbf{A}_S\}, \operatorname{Span}\{\mathbf{B}_{\pi(S)}\}). \quad [25]$$

Since the right-hand side of Eq. (14) is less than one (by Eq. (9) and $r \leq 1$), it follows from Eq. (12) that $|\pi(S)| \geq \dim(\operatorname{Span}\{\mathbf{B}_{\pi(S)}\}) \geq \dim(\operatorname{Span}\{\mathbf{A}_S\}) = |S|$, with the equality by injectivity of $\mathbf{A}$. Since $|S| = |\pi(S)|$, we in fact have $\dim(\operatorname{Span}\{\mathbf{B}_{\pi(S)}\}) = \dim(\operatorname{Span}\{\mathbf{A}_S\})$, and Eq. (25) follows using Eq. (13). Note that the columns of $\mathbf{B}_{\pi(S)}$ are therefore linearly independent, for all $S \in E$.

We next show that Eq. (24) holds. Fix $F \subseteq E$. Since $\operatorname{Span}\{\mathbf{B}_{\cap\pi(F)}\} \subseteq \cap_{S\in F}\operatorname{Span}\{\mathbf{B}_{\pi(S)}\}$, if $\cap_{S\in F}\operatorname{Span}\{\mathbf{B}_{\pi(S)}\} = \{\mathbf{0}\}$, then we must have $|\cap_{S\in F}\pi(S)| = 0$ (as the columns of $\mathbf{B}_{\pi(S)}$ are linearly independent) and Eq. (24) is trivially true. Suppose then that the intersection is not the zero vector. By Lem. 2 and Lem. 3, and then incorporating Eq. (25) and Eq. (14), we have:

$$\begin{aligned}&d(\operatorname{Span}\{\mathbf{B}_{\cap\pi(F)}\}, \operatorname{Span}\{\mathbf{A}_{\cap F}\}) \\ &\quad\leq d\left(\cap_{S\in F}\operatorname{Span}\{\mathbf{B}_{\pi(S)}\}, \cap_{S\in F}\operatorname{Span}\{\mathbf{A}_S\}\right) \\ &\quad\leq \sum_{T\in F}\frac{d\left(\cap_{S\in F}\operatorname{Span}\{\mathbf{B}_{\pi(S)}\}, \operatorname{Span}\{\mathbf{A}_T\}\right)}{r(\operatorname{Span}\{\mathbf{A}_S\}_{S\in F})} \\ &\quad\leq \sum_{T\in F}\frac{d\left(\operatorname{Span}\{\mathbf{B}_{\pi(T)}\}, \operatorname{Span}\{\mathbf{A}_T\}\right)}{r(\operatorname{Span}\{\mathbf{A}_S\}_{S\in F})} \\ &\quad\leq \frac{|F|\varepsilon}{r(\operatorname{Span}\{\mathbf{A}_S\}_{S\in F})} \leq \frac{C_2\varepsilon}{\max_i\|\mathbf{A}_i\|_2}. \quad [26]\end{aligned}$$

where the third inequality follows from the definition of d as a supremum. Since $\varepsilon < L_2(\mathbf{A})/C_2$, by Eq. (9) the right-hand side in Eq. (26) is strictly less than one. Hence, $\dim(\operatorname{Span}\{\mathbf{B}_{\cap\pi(F)}\}) \leq$

$\dim(\mathrm{Span}\{\mathbf{A}_{\cap F}\})$ by Eq. (12), and Eq. (24) follows from the linear independence of the columns of $\mathbf{A}_S$ and $\mathbf{B}_{\pi(S)}$ for all $S \in F$.

Now, by Lem. 4, the association $i \mapsto \cap\pi(F(i))$ defines an injective map $\bar{\pi} : J \to [\bar{m}]$ for some $J \in \binom{[m]}{p}$ with p given by Eq. (23), 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 $S \in E$. It then follows from Eq. (13) and Eq. (26) that $d\left(\mathrm{Span}\{\mathbf{A}_i\}, \mathrm{Span}\{\mathbf{B}_{\bar{\pi}(i)}\}\right) \leq C_2\varepsilon / \max_i \|\mathbf{A}_i\|_2$ 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 Eq. (10), and the result follows from the case $k = 1$ in Sec. 2 applied to the submatrix $\mathbf{A}_J$. □

Proof (sketch) of Thm. 2. Let M be the matrix with columns $\mathbf{A}\mathbf{x}_i$, $i \in [N]$. Consider the polynomial in the entries of $\mathbf{A}$ and $\mathbf{x}_i$:

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

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 $\mathbf{A}$). □

Proof (sketch) of Cor. 3. 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 $\mathbb{R}^p$ (e.g., (37)). By Thm. 2, there is a polynomial that is nonzero whenever Y has a stable k-sparse representation in $\mathbb{R}^m$; in particular, this property (stability) holds with probability one. □