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

by

Charles Garfinkle

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Neuroscience

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Friedrich Sommer, Chair
Bruno Olshausen
Mike Deweese
Bin Yu

Summer 2020

The dissertation of Charles Garfinkle, titled On the uniqueness and stability of dictionaries for sparse representation of noisy signals, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

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

Abstract

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

by

Charles Garfinkle

Doctor of Philosophy in Neuroscience

University of California, Berkeley

Friedrich Sommer, Chair

Learning optimal dictionaries for sparse coding has exposed characteristic sparse features of many natural signals. However, universal guarantees of the stability of such features in the presence of noise are lacking. Here, we provide very general conditions guaranteeing when dictionaries yielding the sparsest encodings are unique and stable with respect to measurement or modeling error. We demonstrate that some or all original dictionary elements are recoverable from noisy data even if the dictionary fails to satisfy the spark condition, its size is overestimated, or only a polynomial number of distinct sparse supports appear in the data. Importantly, we derive these guarantees without requiring any constraints on the recovered dictionary beyond a natural upper bound on its size. Our results also yield an effective procedure sufficient to affirm if a proposed solution to the dictionary learning problem is unique within bounds commensurate with the noise. We suggest applications to data analysis, engineering, and neuroscience and close with some remaining challenges left open by our work.

Dedication

Contents

List of Figures

Acknowledgments

I would like to thank Chris Hillar for laying down the groundwork and vetting all my proofs. He gave me a lot of his time for little if anything in return. Fritz Sommer for taking on the role of my advisor. The Berkeley Student Cooperative for a really awesome and transformative time. CZ or Die. We thank Darren Rhea for early thoughts, and Ian Morris for posting (14) online.

Part I

Main Text

0.1 Introduction

- Background to the topic.
- Brief review of current knowledge.
- Indicate the gap in knowledge. State the aim of the research and how it fits into the gap.
- Can include an outline of what follows.

Sparse coding is a common modern approach to pattern analysis in signal processing whereby each of N observed n-dimensional signal samples is viewed as a (noisy) linear combination of at most k elementary waveforms drawn from some unknown "dictionary" of size $m \ll N$ (see [42] for a comprehensive review). Optimizing dictionaries subject to this and related sparsity constraints has revealed seemingly characteristic sparse structure in several signal classes of current interest (e.g., in vision [40]).

Of particular note are the seminal works in the field [31, 21, 4, 18], which discovered that dictionaries optimized for coding small patches of "natural" images share qualitative similarities with linear filters estimated from response properties of simple-cell neurons in mammalian visual cortex. Curiously, these waveforms (e.g., "Gabor" wavelets) appear in dictionaries learned by a variety of algorithms trained over different natural image datasets, suggesting that learned features in natural signals may, in some sense, be canonical [7].

Motivated by these discoveries and more recent work relating compressed sensing [10] to a theory of information transmission through random wiring bottlenecks in the brain [23], we address when dictionaries for sparse representation are indeed identifiable from data. Answers to this question may also have implications in practice wherever an appeal is made to latent sparse structure of data (e.g., forgery detection [20, 32]; brain recordings [24, 1, 27]; and gene expression [41]).

While several algorithms have been recently proposed to provably recover unique dictionaries under specific conditions (see [36, Sec. I-E] for a summary of the state-of-the-art), few theorems can be invoked to justify the consistency of inference under this model of data more broadly. To our knowledge, a universal guarantee of the uniqueness and stability of learned dictionaries and the sparse representations they induce over noisy data has yet to appear in the literature.

Here, we prove very generally that uniqueness and stability is a typical property of sparse dictionary learning. More specifically, we show that matrices injective on a sparse domain are identifiable from $N = m(k-1)\binom{m}{k} + m$ noisy linear combinations of k of their m 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 problem is well-posed, as per Hadamard [17], given a sufficient amount of data (Thm. 3 and Cor. 2).

Our guarantees also hold for a related (and perhaps more commonly posed, e.g. [34]) optimization problem seeking a dictionary minimizing the average number of elementary

waveforms required to reconstruct each sample of the dataset (Thm. 2). To practical benefit, our results impose no restrictions on learned dictionaries (e.g., that they, too, be injective over some sparse domain) beyond an upper bound on dictionary size, which is necessary in any case to avoid trivial solutions (e.g., allowing $m = N$).

More precisely, 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:

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

for *k-sparse* $\mathbf{x}_i \in \mathbb{R}^m$ having at most $k < m$ nonzero entries and *noise* $\mathbf{n}_i \in \mathbb{R}^n$, with bounded norm $\|\mathbf{n}_i\|_2 \leq \eta$ representing our worst-case uncertainty in measuring the product $\mathbf{A}\mathbf{x}_i$. We first consider the following formulation of sparse coding.

Problem 1. *Find a dictionary matrix* $\mathbf{B}$ *and k-sparse codes* $\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_N$ *that satisfy* $\|\mathbf{z}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \eta$ *for all* $i = 1, \ldots, N$.

Note that every solution to Prob. 1 represents infinitely many equivalent alternatives $\mathbf{BPD}$ and $\mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_1, \ldots, \mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_N$ parametrized by a choice of permutation matrix $\mathbf{P}$ and invertible diagonal matrix $\mathbf{D}$. Identifying these ambiguities (labelling and scale) yields a single orbit of solutions represented by any particular set of elementary waveforms (the columns of $\mathbf{B}$) and their associated sparse coefficients (the entries of $\overline{\mathbf{x}}_i$) that reconstruct each data point $\mathbf{z}_i$.

Previous theoretical work addressing the noiseless case $\eta = 0$ (e.g., [28, 14, 2, 19]) for matrices $\mathbf{B}$ having exactly m columns has shown that a solution to Prob. 1, when it exists, is unique up to such relabeling and rescaling provided the $\mathbf{x}_i$ are sufficiently diverse and $\mathbf{A}$ satisfies the *spark condition*:

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

which is necessary to guarantee the uniqueness of arbitrary k-sparse $\mathbf{x}_i$. We generalize these results to the practical setting $\eta > 0$ by considering the following natural notion of stability with respect to measurement error.

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 some* $\varepsilon = \varepsilon(\delta_1, \delta_2)$ *that is strictly positive for positive* δ_1 *and* δ_2 *such that if* $\mathbf{B}$ *and k-sparse* $\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_N \in \mathbb{R}^m$ *satisfy:*

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

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

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

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

The main result of this work is a very general uniqueness theorem for sparse coding (Thm. 1) directly implying (Cor. 1), which guarantees that sparse representations of a dataset Z are unique up to noise whenever generating dictionaries $\mathbf{A}$ satisfy a spark condition on supports and the original sparse codes $\mathbf{x}_i$ are sufficiently diverse (e.g., Fig. 0.1). Moreover, we provide an explicit, computable $\varepsilon(\delta_1, \delta_2)$ in (8) that is linear in desired accuracy δ_1, and essentially so in δ_2.

In the next section, we give formal statements of these findings. We then extend the same guarantees (Thm. 2) to the following alternate formulation of dictionary learning, which minimizes the total number of nonzero entries in sparse codes.

Problem 2. *Find matrices* $\mathbf{B}$ *and vectors* $\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_N$ *solving:*

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

Our development of Thm. 1 is general enough to provide some uniqueness and stability even when generating $\mathbf{A}$ do not fully satisfy (2) and recovery dictionaries $\mathbf{B}$ have more columns than $\mathbf{A}$. Moreover, the approach incorporates a combinatorial theory for designing generating codes that should be of independent interest. We also give brief arguments adapting our results to dictionaries and codes drawn from probability distributions (Cor. 2). The technical proofs of Thms. 1 and 2 are deferred to Sec. 0.3, following some necessary definitions and a fact in combinatorial matrix analysis (Lem. 1; proven in the Appendix). Finally, we discuss in Sec. 0.4 applications of our mathematical observations as well as open questions.

[1] We mention that the different norms in (3) reflect the distinct meanings typically ascribed to the dictionary and sparse codes in modeling data.

0.2 Results

Precise statements of our results require that we first identify some combinatorial criteria on the supports[2] of sparse vectors. Let $\{1,\dots,m\}$ be denoted $[m]$, its power set $2^{[m]}$, and $\binom{[m]}{k}$ the set of subsets of $[m]$ of size k. A *hypergraph* on vertices $[m]$ is simply any subset $\mathcal{H} \subseteq 2^{[m]}$. We say that $\mathcal{H}$ 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 $2\mathcal{H} := \{S \cup S' : S, S' \in \mathcal{H}\}$. The following class of structured hypergraphs is a key ingredient in this work.

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 $i \in [m]$.*

We next give a quantitative generalization of the spark condition (2) to combinatorial subsets of supports. The *lower bound* of an $n \times m$ matrix $\mathbf{M}$ is the largest α with $\|\mathbf{M}\mathbf{x}\|_2 \geq \alpha\|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{R}^m$ [16]. By compactness of the unit sphere, every injective linear map has a positive lower bound; hence, if $\mathbf{M}$ satisfies (2), then submatrices formed from $2k$ of its columns or less have strictly positive lower bounds.

The lower bound of a matrix is generalized below in (5) by restricting it to the spans of certain submatrices[3] associated with a hypergraph $\mathcal{H} \subseteq \binom{[m]}{k}$ of column indices. Let $\mathbf{M}_S$ denote the submatrix formed by the columns of a matrix $\mathbf{M}$ indexed by $S \subseteq [m]$ (setting $\mathbf{M}_\emptyset := \mathbf{0}$). In the sections that follow, we shall also let $\boldsymbol{\mathcal{M}}_S$ denote the column-span of a submatrix $\mathbf{M}_S$, and $\boldsymbol{\mathcal{M}}_{\mathcal{G}}$ to denote $\{\boldsymbol{\mathcal{M}}_S\}_{S\in\mathcal{G}}$. We define:

$$L_{\mathcal{H}}(\mathbf{M}) := \min\left\{\frac{\|\mathbf{M}_S\mathbf{x}\|_2}{\sqrt{k}\|\mathbf{x}\|_2} : S \in \mathcal{H},\ \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^{|S|}\right\}, \tag{5}$$

writing also L_k in place of $L_{\mathcal{H}}$ when $\mathcal{H} = \binom{[m]}{k}$.[4] As explained above, compactness implies that $L_{2k}(\mathbf{M}) > 0$ for all $\mathbf{M}$ satisfying (2). Clearly, $L_{\mathcal{H}'}(\mathbf{M}) \geq L_{\mathcal{H}}(\mathbf{M})$ whenever $\mathcal{H}' \subseteq \mathcal{H}$, and similarly any k-uniform $\mathcal{H}$ satisfying $\cup\mathcal{H} = [m]$ has $L_2 \geq L_{2\mathcal{H}} \geq L_{2k}$ (letting $L_{2k} := L_m$ whenever $2k > m$).

We are now in a position to state our main result, though for expository purposes we leave the quantity C_1 undefined until Sec. 0.3. All results below assume real matrices and vectors.

Theorem 1. *If an $n \times m$ matrix $\mathbf{A}$ satisfies $L_{2\mathcal{H}}(\mathbf{A}) > 0$ for some r-regular $\mathcal{H} \subseteq \binom{[m]}{k}$ with the SIP, and k-sparse $\mathbf{x}_1,\dots,\mathbf{x}_N \in \mathbb{R}^m$ include more than $(k-1)\binom{m}{k}$ vectors in general linear*

[2]Recall that a vector $\mathbf{x}$ is said to be *supported* in S when $\mathbf{x} \in \operatorname{span}\{\mathbf{e}_j : j \in S\}$, with $\mathbf{e}_j$ forming the standard column basis.

[3]See [39] for an overview of the related "union of subspaces" model.

[4]In compressed sensing literature, $1 - \sqrt{k}L_k(\mathbf{M})$ is the asymmetric lower restricted isometry constant for $\mathbf{M}$ with unit ℓ_2-norm columns [5].

position[5] *supported in each* $S \in \mathcal{H}$*, then the following recovery guarantees hold for* $C_1 > 0$ *given by* (19).

Dictionary Recovery: *Fix* $\varepsilon < L_2(\mathbf{A})/C_1$.[6] *If an* $n \times \overline{m}$ *matrix* $\mathbf{B}$ *has, for every* $i \in [N]$*, an associated* k*-sparse* $\overline{\mathbf{x}}_i$ *satisfying* $\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \varepsilon$*, then* $\overline{m} \geq m$*, and provided that* $\overline{m}(r-1) < mr$*, there is a permutation matrix* $\mathbf{P}$ *and an invertible diagonal matrix* $\mathbf{D}$ *such that:*

$$\|\mathbf{A}_j - (\mathbf{BPD})_j\|_2 \leq C_1\varepsilon, \quad \text{for all } j \in J, \tag{6}$$

for some $J \subseteq [m]$ *of size* $m - (r-1)(\overline{m} - m)$.

Code Recovery: *If, moreover,* $\mathbf{A}_J$ *satisfies* (2) *and* $\varepsilon < L_{2k}(\mathbf{A}_J)/C_1$*, then* $(\mathbf{BP})_J$ *also satisfies* (2) *with* $L_{2k}(\mathbf{BP}_J) \geq (L_{2k}(\mathbf{A}_J) - C_1\varepsilon)/\|\mathbf{D}_J\|_1$*, and for all* $i \in [N]$*:*

$$\|(\mathbf{x}_i)_J - (\mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_i)_J\|_1 \leq \left(\frac{1 + C_1\|(\mathbf{x}_i)_J\|_1}{L_{2k}(\mathbf{A}_J) - C_1\varepsilon}\right)\varepsilon, \tag{7}$$

where subscript $(\cdot)_J$ *here represents the subvector formed from restricting to coordinates indexed by* J.

In words, Thm. 1 says that the smaller the regularity r of the original support hypergraph $\mathcal{H}$ or the difference $\overline{m} - m$ between the assumed and actual number of elements in the latent dictionary, the more columns and coefficients of the original dictionary $\mathbf{A}$ and sparse codes $\mathbf{x}_i$ are guaranteed to be contained (up to noise) in the appropriately labelled and scaled recovered dictionary $\mathbf{B}$ and codes $\overline{\mathbf{x}}_i$, respectively.

In the important special case when $\overline{m} = m$, the theorem directly implies that $Y = \{\mathbf{A}\mathbf{x}_1, \ldots, \mathbf{A}\mathbf{x}_N\}$ has a stable k-sparse representation in $\mathbb{R}^m$, with inequalities (3) guaranteed in Def. 1 for the following worst-case error ε:

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

Since sparse codes in general linear position are straightforward to produce with a "Vandermonde" construction (i.e., by choosing columns of the matrix $[\gamma_i^j]_{i,j=1}^{k,N}$, for distinct nonzero γ_i), we have the following direct consequence of Thm. 1.

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

[5] Recall that a set of vectors sharing support S are in *general linear position* when any $|S|$ of them are linearly independent.

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

One can easily verify that for every $k < m$ there are regular k-uniform hypergraphs $\mathcal{H}$ with the SIP besides the obvious $\mathcal{H} = \binom{[m]}{k}$. For instance, take $\mathcal{H}$ to be the k-regular set of 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. Special cases allow for even smaller hypergraphs. For example, if $k = \sqrt{m}$, then a 2-regular k-uniform hypergraph with the SIP can be constructed as the $2k$ rows and columns formed by arranging the elements of $[m]$ into a square grid.

We should stress here that framing the problem in terms of hypergraphs has allowed us to show, unlike in previous research on the subject, that the matrix $\mathbf{A}$ need not necessarily satisfy (2) to be recoverable from data. As an example, let $\mathbf{A} = [\mathbf{e}_1, \ldots, \mathbf{e}_5, \mathbf{v}]$ with $\mathbf{v} = \mathbf{e}_1 + \mathbf{e}_3 + \mathbf{e}_5$ and take $\mathcal{H}$ to be all consecutive pairs of indices $1, \ldots, 6$ arranged in cyclic order. Then for $k = 2$, the matrix $\mathbf{A}$ fails to satisfy (2) while still obeying the assumptions of Thm. 1 for dictionary recovery.

A practical implication of Thm. 1 is the following: there is an effective procedure sufficient to affirm if a proposed solution to Prob. 1 is indeed unique (up to noise and inherent ambiguities). One need simply check that the matrix and codes satisfy the (computable) assumptions of Thm. 1 on $\mathbf{A}$ and the $\mathbf{x}_i$. In general, however, there is no known efficient procedure. We defer a brief discussion on this point to the next section.

A less direct consequence of Thm. 1 is the following uniqueness and stability guarantee for solutions to Prob. 2.

Theorem 2. *Fix a matrix* $\mathbf{A}$ *and vectors* $\mathbf{x}_i$ *satisfying the assumptions of Thm. 1, only now with over* $(k-1)\left[\binom{\overline{m}}{k} + |\mathcal{H}|k\binom{\overline{m}}{k-1}\right]$ *vectors supported in general linear position in each* $S \in \mathcal{H}$*. Every solution to Prob. 2 (with* $\eta = \varepsilon/2$*) satisfies recovery guarantees* (6) *and* (7) *when the corresponding bounds on* η *are met.*

Another extension of Thm. 1 can be derived from the following algebraic characterization of the spark condition. Letting $\mathbf{A}$ be the $n \times m$ matrix of nm indeterminates A_{ij}, the reader may work out why substituting real numbers for the A_{ij} yields a matrix satisfying (2) if and only if the following polynomial evaluates to a nonzero number:

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

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 $(i,j) \in S' \times S$.[7]

Since f is analytic, having a single substitution of a real matrix $\mathbf{A}$ satisfying $f(\mathbf{A}) \neq 0$ implies that the zeroes of f form a set of (Borel) measure zero. Such a matrix is easily constructed by adding rows of zeroes to a $\min(2k, m) \times m$ Vandermonde matrix as mentioned previously, so that every sum in the product defining f above is strictly positive. Thus, almost every $n \times m$ matrix with $n \geq \min(2k, m)$ satisfies (2).

[7]The large number of terms in this product is likely necessary given that deciding whether or not a matrix satisfies the spark condition is NP-hard [38].

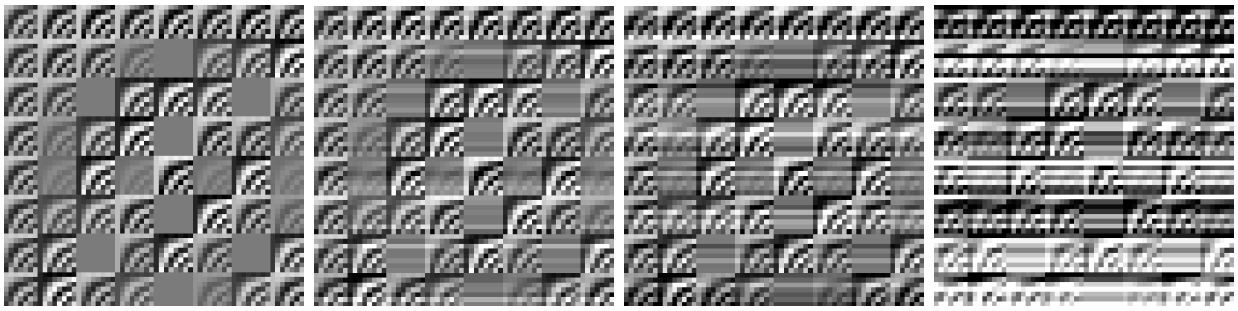

Figure 0.1: **Learning a dictionary from increasingly noisy data**. The (unraveled) basis elements of the 8×8 discrete cosine transform (DCT) form the 64 columns of the left-most matrix above. Three increasingly imprecise dictionaries (columns reordered to best match original) are recovered by FastICA [22] trained on data generated from 8-sparse linear combinations of DCT elements corrupted with additive noise (increasing from left to right).

We claim that a similar phenomenon applies to datasets of vectors with a stable sparse representation. Briefly, following the same procedure as in [19, Sec. IV], for $k < m$ and $n \geq \min(2k, m)$, we may consider the "symbolic" dataset $Y = \{\mathbf{A}\mathbf{x}_1, \ldots, \mathbf{A}\mathbf{x}_N\}$ generated by an indeterminate $n \times m$ matrix $\mathbf{A}$ and m-dimensional k-sparse vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ indeterminate within their supports, which form a regular hypergraph $\mathcal{H} \subseteq \binom{[m]}{k}$ satisfying the SIP. Restricting $(k-1)\binom{m}{k} + 1$ indeterminate $\mathbf{x}_i$ to each support in $\mathcal{H}$, and letting $\mathbf{M}$ be the $n \times N$ matrix with columns $\mathbf{A}\mathbf{x}_i$, it can be checked that when $f(\mathbf{M}) \neq 0$ for a substitution of real numbers for the indeterminates, all of the assumptions on $\mathbf{A}$ and the $\mathbf{x}_i$ in Thm. 1 are satisfied. We therefore have the following.

Theorem 3. *There is a polynomial in the entries of* $\mathbf{A}$ *and the* $\mathbf{x}_i$ *that evaluates to a nonzero number only when* Y *has a stable* k*-sparse representation in* $\mathbb{R}^m$*. In particular, almost all substitutions impart to* Y *this property.*

To extend this observation to arbitrary probability distributions, note that if a set of p measure spaces has all measures absolutely continuous with respect to the standard Borel measure on $\mathbb{R}$, then the product measure is also absolutely continuous with respect to the standard Borel product measure on $\mathbb{R}^p$ (e.g., see [11]). This fact combined with Thm. 3 implies the following.[8]

Corollary 2. *If the indeterminate entries of* $\mathbf{A}$ *and the* $\mathbf{x}_i$ *are drawn independently from probability distributions absolutely continuous with respect to the standard Borel measure, then* Y *has a stable* k*-sparse representation in* $\mathbb{R}^m$ *with probability one.*

[8] We refer the reader to [19] for a more detailed explanation of these arguments.

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

We close this section with some comments on the optimality of our results. The linear scaling for ε in (8) is essentially optimal (e.g., see [3]), but a basic open problem remains: how many samples are necessary to determine the sparse coding model? Our results demonstrate that sparse codes $\mathbf{x}_i$ drawn from only a polynomial number of k-dimensional subspaces permit stable identification of the generating dictionary $\mathbf{A}$. This lends some legitimacy to the use of the model in practice, where data in general are unlikely (if ever) to exhibit the exponentially many possible k-wise combinations of dictionary elements required by (to our knowledge) all previously published results.

Consequently, if k is held fixed or if the size of the support set of reconstructing codes is polynomial in $\overline{m}$ and k, then a practical (polynomial) amount of data suffices to identify the dictionary.[9] Reasons to be skeptical that this holds in general, however, can be found in [38, 37]. Even so, in the next section we discuss how probabilistic guarantees can in fact be made for any number of available samples.

[9]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.

0.3 Proofs

We begin our proof of Thm. 1 by showing how dictionary recovery (6) already implies sparse code recovery (7) when $\mathbf{A}$ satisfies (2) and $\varepsilon < L_{2k}(\mathbf{A})/C_1$. We temporarily assume (without loss of generality) that $\overline{m} = m$, so as to omit an otherwise requisite subscript $(\cdot)_J$ around certain matrices and vectors. By definition of L_{2k} in (5), and noting that $\sqrt{k}\|\mathbf{v}\|_2 \geq \|\mathbf{v}\|_1$ for k-sparse $\mathbf{v}$, we have for all $i \in [N[$:

$$\begin{aligned}\|\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top \overline{\mathbf{x}}_i\|_1 &\leq \frac{\|\mathbf{BPD}(\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top \overline{\mathbf{x}}_i)\|_2}{L_{2k}(\mathbf{BPD})} \\ &\leq \frac{\|(\mathbf{BPD} - \mathbf{A})\mathbf{x}_i\|_2 + \|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2}{L_{2k}(\mathbf{BPD})} \\ &\leq \frac{C_1\varepsilon\|\mathbf{x}_i\|_1 + \varepsilon}{L_{2k}(\mathbf{BPD})}, \end{aligned} \tag{9}$$

where the first term in the numerator above follows from the triangle inequality and (6).

It remains for us to bound the denominator. For any $2k$-sparse $\mathbf{x}$, we have by the triangle inequality:

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

We therefore have that $L_{2k}(\mathbf{BPD}) \geq L_{2k}(\mathbf{A}) - C_1\varepsilon > 0$, and (7) then follows from (9). The reader may also verify that $L_{2k}(\mathbf{BP}) \geq L_{2k}(\mathbf{BPD})/\|\mathbf{D}\|_1$.

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

Proof of Thm. 1 for $k = 1$. Since the only 1-uniform hypergraph with the SIP is $[m]$, which is obviously regular, we require only $\mathbf{x}_i = c_i\mathbf{e}_i$ for $i \in [m]$, with $c_i \neq 0$ to guarantee linear independence. While we have yet to define C_1 generally, in this case we may set $C_1 = 1/\min_{\ell\in[m]}|c_\ell|$ so that $\varepsilon < L_2(\mathbf{A})\min_{\ell\in[m]}|c_\ell|$.

Fix $\mathbf{A} \in \mathbb{R}^{n\times m}$ satisfying $L_2(\mathbf{A}) > 0$, since here we have $2\mathcal{H} = \binom{[m]}{2}$, and suppose some $\mathbf{B}$ and 1-sparse $\overline{\mathbf{x}}_i \in \mathbb{R}^{\overline{m}}$ have $\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \varepsilon < L_2(\mathbf{A})/C_1$ for all i. Then, there exist $\overline{c}_1, \ldots, \overline{c}_m \in \mathbb{R}$ and a map $\pi : [m] \to [\overline{m}]$ such that:

$$\|c_i\mathbf{A}_i - \overline{c}_i\mathbf{B}_{\pi(i)}\|_2 \leq \varepsilon, \quad \text{for } i \in [m]. \tag{10}$$

Note that $\overline{c}_i \neq 0$, since otherwise we would reach the following contradiction: $\|\mathbf{A}_i\|_2 \leq C_1|c_i|\|\mathbf{A}_i\|_2 \leq C_1\varepsilon < L_2(\mathbf{A}) \leq L_1(\mathbf{A}) = \min_{i\in[m]}\|\mathbf{A}_i\|_2$.

We now show that π is injective (in particular, a permutation if $\overline{m} = m$). Suppose that $\pi(i) = \pi(j) = \ell$ for some $i \neq j$ and ℓ. Then, $\|c_i\mathbf{A}_i - \overline{c}_i\mathbf{B}_\ell\|_2 \leq \varepsilon$ and $\|c_j\mathbf{A}_j - \overline{c}_j\mathbf{B}_\ell\|_2 \leq \varepsilon$,

and we have:

$$\begin{aligned}(|\overline{c}_i| + |\overline{c}_j|)\varepsilon &\geq |\overline{c}_i|\|c_j\mathbf{A}_j - \overline{c}_j\mathbf{B}_\ell\|_2 + |\overline{c}_j|\|c_i\mathbf{A}_i - \overline{c}_i\mathbf{B}_\ell\|_2 \\ &\geq \|\mathbf{A}(\overline{c}_i c_j \mathbf{e}_j - \overline{c}_j c_i \mathbf{e}_i)\|_2 \\ &\geq \sqrt{2}L_2(\mathbf{A})\|\overline{c}_i c_j \mathbf{e}_j - \overline{c}_j c_i \mathbf{e}_i\|_2 \\ &\geq L_2(\mathbf{A})\,(|\overline{c}_i| + |\overline{c}_j|)\min_{\ell\in[m]}|c_\ell|,\end{aligned}$$

contradicting our assumed upper bound on ε. Hence, the map π is injective and so $\overline{m} \geq m$.

Letting $\mathbf{P}$ and $\mathbf{D}$ be the $\overline{m} \times \overline{m}$ permutation and invertible diagonal matrices with, respectively, columns $\mathbf{e}_{\pi(i)}$ and $\frac{\overline{c}_i}{c_i}\mathbf{e}_i$ for $i \in [m]$ (otherwise, $\mathbf{e}_i$ for $i \in [\overline{m}] \setminus [m]$), we may rewrite (10) to see that for all $i \in [m]$:

$$\|\mathbf{A}_i - (\mathbf{BPD})_i\|_2 = \|\mathbf{A}_i - \frac{\overline{c}_i}{c_i}\mathbf{B}_{\pi(i)}\|_2 \leq \frac{\varepsilon}{|c_i|} \leq C_1\varepsilon.$$

□

An extension of the proof to the general case $k < m$ requires some additional tools to derive the general expression (19) for C_1. 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 dist$(\mathbf{u}, V) := \min\{\|\mathbf{u} - \mathbf{v}\|_2 : \mathbf{v} \in V\}$ and define:*

$$d(U, V) := \max_{\mathbf{u}\in U,\ \|\mathbf{u}\|_2\leq 1} \mathit{dist}(\mathbf{u}, V). \tag{11}$$

We note the following facts about d. Clearly,

$$U' \subseteq U \implies d(U', V) \leq d(U, V). \tag{12}$$

From [25, Ch. 4 Cor. 2.6], we also have:

$$d(U, V) < 1 \implies \dim(U) \leq \dim(V), \tag{13}$$

and from [30, Lem. 3.2]:

$$\dim(U) = \dim(V) \implies d(U, V) = d(V, U). \tag{14}$$

The following is our result in combinatorial matrix analysis; it contains most of the complexity in the proof of Thm. 1.

Lemma 1. *If an $n \times m$ matrix $\mathbf{A}$ has $L_{2\mathcal{H}}(\mathbf{A}) > 0$ for some r-regular $\mathcal{H} \subseteq \binom{[m]}{k}$ with the SIP, then the following holds for $C_2 > 0$ given by (18):*

Fix $\varepsilon < L_2(\mathbf{A})/C_2$. If for some $n \times \overline{m}$ matrix $\mathbf{B}$ and map $\pi : \mathcal{H} \mapsto \binom{[\overline{m}]}{k}$,

$$d(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\pi(S)}) \leq \varepsilon, \quad \text{for } S \in \mathcal{H}, \tag{15}$$

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

$$\|\mathbf{A}_i - (\mathbf{BPD})_i\|_2 \leq C_2\varepsilon, \quad \text{for } i \in J, \tag{16}$$

for some $J \subseteq [m]$ of size $m - (r-1)(\overline{m} - m)$.

We present the constant C_2 (a function of $\mathbf{A}$ and $\mathcal{H}$) relative to a quantity used in [6] to analyze the convergence of the "alternating projections" algorithm for projecting a point onto the intersection of subspaces. We incorporate this quantity into the following definition, which we refer to in our proof of Lem. 3 in the Appendix; specifically, we use it to bound the distance between a point and the intersection of subspaces given an upper bound on its distance from each subspace.

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

$$\xi^2(\mathcal{V}) := 1 - \max \prod_{i=1}^{\ell-1} \sin^2\theta\left(V_i, \cap_{j>i} V_j\right), \tag{17}$$

where the maximum is taken over all ways of ordering the V_i and the angle $\theta \in (0, \frac{\pi}{2}]$ is defined implicitly as [6, 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 $\theta \in (0, \frac{\pi}{2}]$ implies $0 \leq \xi < 1$, and that $\xi(\mathcal{V}') \leq \xi(\mathcal{V})$ when $\mathcal{V}' \subseteq \mathcal{V}$.[10]

The constant $C_2 > 0$ of Lem. 1 can now be expressed as:

$$C_2(\mathbf{A}, \mathcal{H}) := \frac{(r+1)\max_{j\in[m]} \|\mathbf{A}_j\|_2}{1 - \max_{\mathcal{G} \in \binom{\mathcal{H}}{r+1}} \xi(\boldsymbol{\mathcal{A}}_{\mathcal{G}})}. \tag{18}$$

We next define the constant $C_1 > 0$ of Thm. 1 in terms of C_2. Given vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$, let $\mathbf{X}$ denote the $m \times N$ matrix with columns $\mathbf{x}_i$ and let $I(S)$ denote the set of indices i for which $\mathbf{x}_i$ is supported in S. We define:

$$C_1(\mathbf{A}, \mathcal{H}, \{\mathbf{x}_i\}_{i=1}^N) := \frac{C_2(\mathbf{A}, \mathcal{H})}{\min_{S\in\mathcal{H}} L_k(\mathbf{A}\mathbf{X}_{I(S)})}. \tag{19}$$

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

Given the assumptions of Thm. 1 on $\mathbf{A}$ and the $\mathbf{x}_i$, this expression for C_1 is well-defined[11] and yields an upper bound on ε consistent with that proven sufficient in the case $k = 1$ considered at the beginning of this section.[12]

The practically-minded reader should note that the explicit constants C_1 and C_2 are effectively computable: the denominator of C_1 involves a quantity L_k that may be calculated as the smallest singular value of a certain matrix, while computing the quantity ξ in the denominator of C_2 involves computing "canonical angles" between subspaces, which reduces 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 [38], although efficiently computable bounds have been proposed (e.g., via the "mutual coherence" of a matrix [8]); alternatively, fixing k yields polynomial complexity. Moreover, calculating C_2 requires an exponential number of queries to ξ unless r is held fixed, too (e.g., the "cyclic order" hypergraphs described above have $r = k$). Thus, as presented, C_1 and C_2 are not efficiently computable in general.

We note also that these constants likely have reasonable values in general. The quantity L_k is a relatively standard quantity in the field of compressed sensing (the "restricted isometry constant") and it is known to be reasonable for many random distributions generating dictionaries $\mathbf{A}$ and sparse codes $\mathbf{x}_i$. The more obscure quantity ξ, on the other hand, is computed from the "Friedrichs angle" between certain spans of subsets of the columns of $\mathbf{A}$. Simulations for small random matrices $\mathbf{A}$ and hypergraphs $\mathcal{H}$ satisfying the SIP suggest nonetheless that the constant C_2 is reasonable as well (Fig. 0.5).

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

Fix $S \in \mathcal{H}$. Since there are more than $(k-1)\binom{\overline{m}}{k}$ vectors $\mathbf{x}_i$ supported in S, by the pigeonhole principle there must be some $\overline{S} \in \binom{[\overline{m}]}{k}$ and a set of k indices $K \subseteq I(S)$ for which all $\overline{\mathbf{x}}_i$ with $i \in K$ are supported in $\overline{S}$. It also follows[13] from $L_{2\mathcal{H}}(\mathbf{A}) > 0$ and the general linear position of the $\mathbf{x}_i$ that $L_k(\mathbf{A}\mathbf{X}_K) > 0$; that is, the columns of the $n \times k$ matrix $\mathbf{A}\mathbf{X}_K$ form a basis for $\boldsymbol{\mathcal{A}}_S$.

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

$$\begin{aligned}\|\mathbf{y} - \overline{\mathbf{y}}\|_2 = \|(\mathbf{A}\mathbf{X}_K - \mathbf{B}\overline{\mathbf{X}}_K)\mathbf{c}\|_2 &\leq \varepsilon\|\mathbf{c}\|_1 \leq \varepsilon\sqrt{k}\|\mathbf{c}\|_2 \\ &\leq \frac{\varepsilon}{L_k(\mathbf{A}\mathbf{X}_K)}\|\mathbf{y}\|_2,\end{aligned}$$

[11] To see this, fix $S \in \mathcal{H}$ and k-sparse $\mathbf{c}$. Using the definitions, we have $\|\mathbf{A}\mathbf{X}_{I(S)}\mathbf{c}\|_2 \geq \sqrt{k}L_{\mathcal{H}}(\mathbf{A})\|\mathbf{X}_{I(S)}\mathbf{c}\|_2 \geq kL_{\mathcal{H}}(\mathbf{A})L_k(\mathbf{X}_{I(S)})\|\mathbf{c}\|_2$. Thus, $L_k(\mathbf{A}\mathbf{X}_{I(S)}) \geq \sqrt{k}L_{\mathcal{H}}(\mathbf{A})L_k(\mathbf{X}_{I(S)}) > 0$, since $L_{\mathcal{H}}(\mathbf{A}) \geq L_{2\mathcal{H}}(\mathbf{A}) > 0$ and $L_k(\mathbf{X}_{I(S)}) > 0$ by general linear position of the $\mathbf{x}_i$.

[12] When $\mathbf{x}_i = c_i\mathbf{e}_i$, we have $C_2 \geq 2\|\mathbf{A}_i\|_2$ and the denominator in (19) becomes $\min_{i\in[m]}|c_i|\|\mathbf{A}_i\|_2$; hence, $C_1 \geq 2/\min_{i\in[m]}|c_i|$.

[13] See footnote 11.

where the last inequality uses (5). From Def. 3:

$$d(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\overline{S}}) \leq \frac{\varepsilon}{L_k(\mathbf{A}\mathbf{X}_K)} \leq \frac{\varepsilon}{L_k(\mathbf{A}\mathbf{X}_{I(S)})} \leq \varepsilon \frac{C_1}{C_2}. \tag{20}$$

Finally, apply Lem. 1 with $\varepsilon < L_2(\mathbf{A})/C_1$ and $\pi(S) := \overline{S}$. □

Before moving on to the proof of Thm. 2, we briefly revisit our discussion on sample complexity from the end of the previous section. While an exponential number of samples may very well prove to be necessary in the deterministic or almost-certain case, our proof of Thm. 1 can be extended to hold with some probability for *any* number of samples by alternative appeal to a probabilistic pigeonholing at the point early in the proof where the (deterministic) pigeonhole principle is applied to show that for every $S \in \mathcal{H}$, there exist k vectors $\mathbf{x}_i$ supported on S whose corresponding $\overline{\mathbf{x}}_i$ all share the same support.[14] Given insufficient samples, this argument has some less-than-certain probability of being valid for each $S \in \mathcal{H}$. Nonetheless, simulations with small hypergraphs demonstrate that successful recovery is nearly certainly even when N is only a fraction of the deterministic sample complexity (Fig. 0.5).

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

$$\sum_{p=0}^{\overline{m}} p n_p = \sum_{i=0}^{N} \|\overline{\mathbf{x}}_i\|_0 \leq \sum_{i=0}^{N} \|\mathbf{x}_i\|_0 \leq kN$$

Since $N = \sum_{p=0}^{\overline{m}} n_p$, we then have $\sum_{p=0}^{\overline{m}} (p-k) n_p \leq 0$. Splitting the sum yields:

$$\sum_{p=k+1}^{\overline{m}} n_p \leq \sum_{p=k+1}^{\overline{m}} (p-k) n_p \leq \sum_{p=0}^{k} (k-p) n_p \leq k \sum_{p=0}^{k-1} n_p, \tag{21}$$

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

Next, observe that no more than $(k-1)|\mathcal{H}|$ of the $\overline{\mathbf{x}}_i$ share a support $\overline{S}$ of size less than k. Otherwise, by the pigeonhole principle, there is some $S \in \mathcal{H}$ and a set of k indices $K \subseteq I(S)$ for which all $\mathbf{x}_i$ with $i \in K$ are supported in S; as argued previously, (20) follows. It is simple to show that $L_2(\mathbf{A}) \leq \max_j \|\mathbf{A}_j\|_2$, and since $0 \leq \xi < 1$, the right-hand side of (20) is less than one for $\varepsilon < L_2(\mathbf{A})/C_1$. Thus, by (13) we would have the contradiction $k = \dim(\boldsymbol{\mathcal{A}}_S) \leq \dim(\boldsymbol{\mathcal{B}}_{\overline{S}}) \leq |\overline{S}| < k$.

The total number of $(k-1)$-sparse vectors $\overline{\mathbf{x}}_i$ thus cannot exceed $|\mathcal{H}|(k-1)\binom{\overline{m}}{k-1}$. By (21), no more than $|\mathcal{H}|k(k-1)\binom{\overline{m}}{k-1}$ vectors $\overline{\mathbf{x}}_i$ are not k-sparse. Since for every $S \in \mathcal{H}$ there

[14] A famous example of such an argument is the counter-intuitive "birthday paradox", which demonstrates that the probability of two people having the same birthday in a room of twenty-three is greater than 50%.

are over $(k-1)\left[\binom{\overline{m}}{k} + |\mathcal{H}|k\binom{\overline{m}}{k-1}\right]$ vectors $\mathbf{x}_i$ supported there, it must be that more than $(k-1)\binom{\overline{m}}{k}$ of them have corresponding $\overline{\mathbf{x}}_i$ that are k-sparse. The result now follows from Thm. 1, noting by the triangle inequality that $\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\| \leq 2\eta$ for $i = 1, \ldots, N$. $\square$

0.4 Discussion

A main motivation for this work is the emergence of seemingly unique representations from sparse coding models trained on natural data, despite the varied assumptions underlying the many algorithms in current use. Our results constitute an important step toward explaining these phenomena as well as unifying many publications on the topic by deriving general deterministic conditions under which identification of parameters in this model is not only possible but also robust to uncertainty in measurement and model choice.

We have shown that, given sufficient data, the problem of determining a dictionary and sparse codes with minimal support size (Prob. 2) reduces to an instance of Prob. 1, to which our main result (Thm. 1) applies: every dictionary and sequence of sparse codes consistent with the data are equivalent up to inherent relabeling/scaling ambiguities and a 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 these problems is indeed unique up to noise and inherent ambiguities, although it is not efficient in general.

Beyond an extension of existing noiseless guarantees [19] to the noisy regime and their novel application to Prob. 2, our work contains a theory of combinatorial designs for support sets key to identification of dictionaries. We incorporate this idea into a fundamental lemma in matrix theory (Lem. 1) that draws upon the definition of a matrix lower bound (5) induced by a hypergraph. The new insight offered by this combinatorial approach allows for guaranteed recovery of some or all dictionary elements even if: 1) dictionary size is overestimated, 2) data cover only a polynomial number of distinct sparse supports, and 3) dictionaries do not satisfy the spark condition.

The absence of any assumptions about dictionaries solving Prob. 1 was a major technical obstruction in proving Thm. 1. We sought such a general guarantee because of the practical difficulty in ensuring that an algorithm maintain a dictionary satisfying the spark condition (2) at each iteration, an implicit requirement of all previous works except [19]; indeed, even certifying a dictionary has this property is NP-hard [38].

One direct application of this work is to theoretical neuroscience, wherein our theorems justify the mathematical soundness of one of the few hypothesized theories of bottleneck communication in the brain [23]: that sparse neural population activity is recoverable from its noisy linear compression through a randomly constructed (but unknown) wiring bottleneck by any biologically plausible unsupervised sparse coding method that solves Prob. 1 or 2 (e.g., [34, 35, 33]).[15]

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; for instance, in MRI analysis [29], image compression [9], and even the design of an ultrafast camera [13]. It is only a matter of time before these systems incorporate dictionary learning to encode and decode signals

[15] We refer the reader to [12] for more on interactions between dictionary learning and neuroscience.

(e.g., in a device that learns structure from motion [26]), just as scientists have used sparse coding to make sense of their data [24, 1, 27, 41].

Assurances offered by our theorems certify that different devices and algorithms learn equivalent representations given enough data from statistically identical systems.[16] Indeed, a main reason for the sustained interest in dictionary learning as an unsupervised method for data analysis seems to be the assumed well-posedness of parameter identification in the model, confirmation of which forms the core of our findings.

0.5 Future Directions

There are many challenges left open by this work. All conditions stated here guaranteeing the uniqueness and stability of sparse representations have only been shown sufficient; it remains open, therefore, to extend them to necessary conditions, be they on required sample size, the structure of support set hypergraphs, or tolerable error bounds. On this last note, we remark that the deterministic conditions derived here must consider always the "worst-case" noise, whereas the "effective" noise sampled from a concentrated distribution might be significantly reduced, especially for high-dimensional data. It would be of great practical benefit to see how drastically all conditions can be relaxed by requiring only probabilistic guarantees in this way, or in the spirit of our discussion on probabilistic pigeonholing to reduce sample complexity (as in the famous "birthday paradox") following the proof of Thm. 1.

Another interesting question raised by this work is for which special cases is it efficient to check that a solution to Prob. 1 or 2 is unique up to noise and inherent ambiguities. Considering that the sufficient conditions we have described for checking this in general are NP-hard to compute, are the necessary conditions hard? Are Probs. 1 and 2 then also hard (e.g., see [37])? Finally, since Prob. 2 is intractable in general, but efficiently solvable by ℓ_1-norm minimization when the matrix is known (and has a large enough lower bound over sparse domains [10]), is there a version of Thm. 2 certifying when Prob. 2 can be solved efficiently in full by similar means?

We hope these remaining challenges pique the interest of the community to pick up where we have left off and that the theoretical tools showcased here may be of use to this end.

[16] 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 [15].

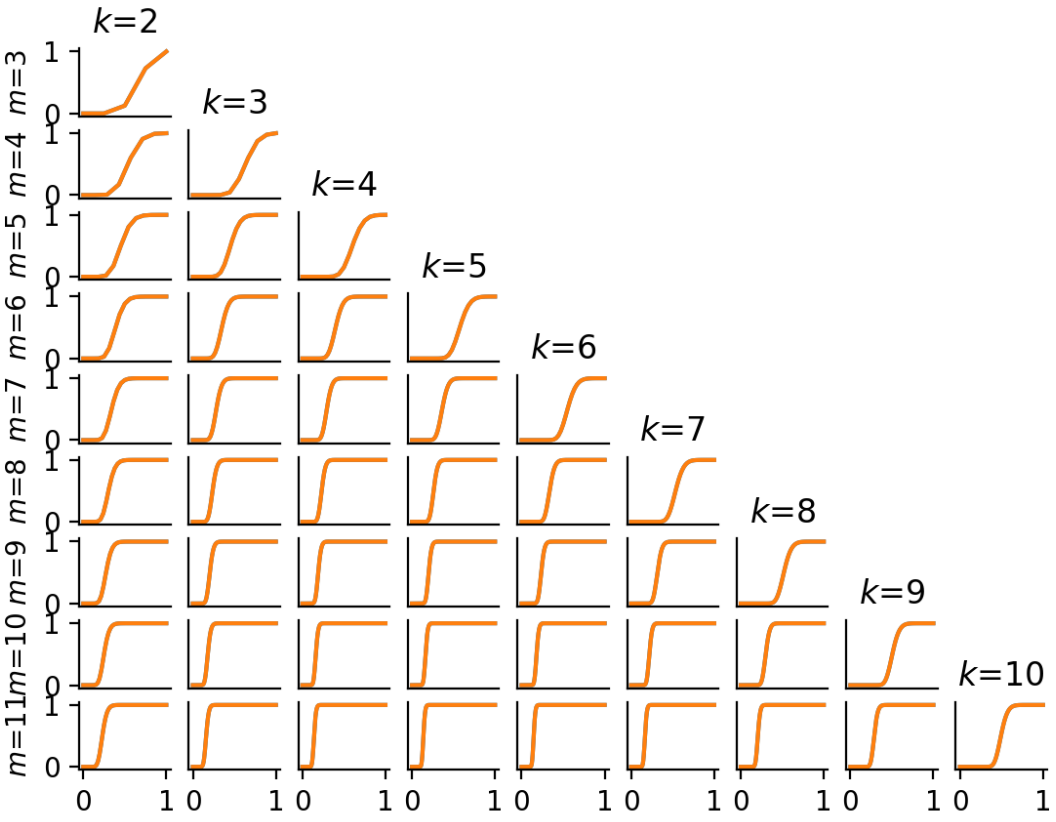

Figure 0.2: Probability of successful dictionary and code recovery (as per Thm. 1) for a number of samples N given as a fraction of the deterministic sample complexity $N = |\mathcal{H}|[(k-1)\binom{m}{k} + 1]$ when $\mathcal{H}$ is taken to be the set of m consecutive intervals of length k in a cyclic order on $[m]$ (i.e. $|\mathcal{H}| = m$). Successful recovery is nearly certain far below the deterministic sample complexity.

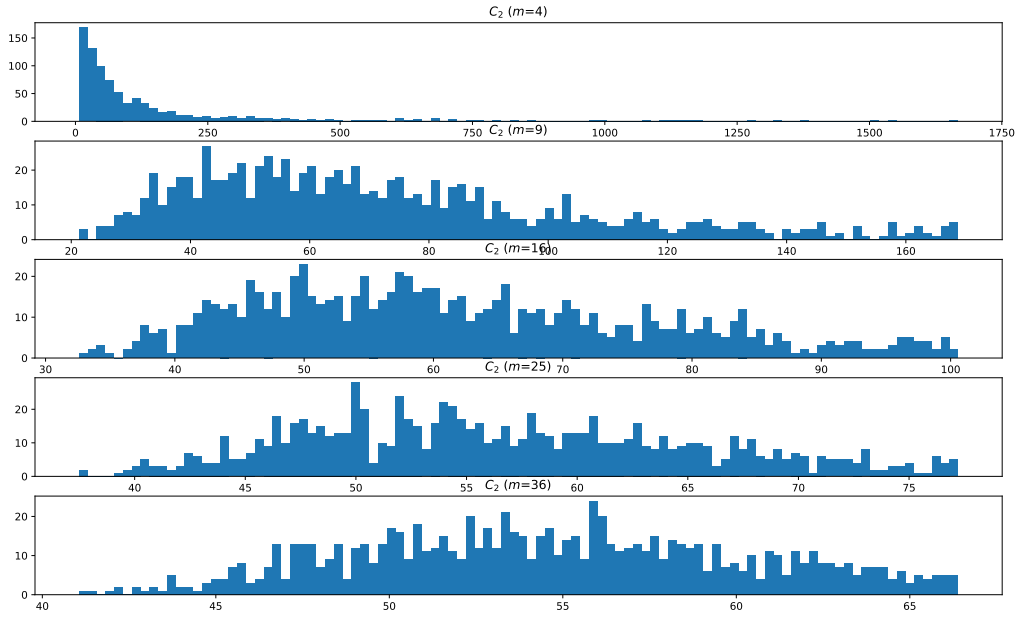

Figure 0.3: The constant $C_2(\mathbf{A}, \mathcal{H})$ computed for generic unit-norm matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$ and the hypergraph $\mathcal{H}$ consisting of the rows and columns formed by arranging the elements of $[m]$ into a square grid. The results suggest that the bound is typically concentrated around a reasonable value.

0.6 Conclusions

HOW DO WE IMPROVE ON HS15?

??????

REVIEWER: Specifically, the authors could address why going from the noise-free case to the noisy case is not incremental. PNAS specifically targets submissions that go beyond what could be published in a more specialized journal. Could the authors, for example, contrast the proof technique between the noisy and noise-free cases? Or, perhaps, the practical implications?

RESPONSE: With all due respect to the reviewer, we believe the most compelling practical implication of a noisy result over a noiseless result is plainly evident: never has there been an experiment without noise and an unstable model is therefore useless in practice. We assume the reviewer must instead be concerned with what additional practical "take-away" message our manuscript contains. One significant one is the outline of a procedure that can (sufficiently) affirm if one?s proposed solution to Prob. 1 or Prob. 2 is indeed unique. We have updated the manuscript with an explicit statement of this fact, and we outline one such procedure here:

Given a dictionary A and codes x-i that solve Problem 1 (e.g., learned by any algorithm), check that that they satisfy the assumptions of Thm. 1: - List the support sets of the x-i. - Discard those for which there are (k-1)(m /choose k) or fewer x-i with that support. - Discard those for which the supported x-i are not in general linear position. (This should also exclude any supports less than k in size.) - From the support sets that remain, list all SIP-satisfying hypergraphs that are subsets of this set. - Discard those for which L-(2H)(A) = 0. - For each of the remaining hypergraphs, determine the constant C-1 from A and the x-i (or for every subset of the x-i of size (k-1)(m /choose k)). - Check that the derived upper bound on /varepsilon exceeds that of the original problem. If yes, the solution is "unique".

There are other practical (polynomial time) implications of our discovery of sufficient combinatorial designs for support sets of generating codes x-i. Moreover, we have shown that a subset of dictionary elements are recoverable even if the number of dictionary elements in total is unknown; these observations are discussed in more detail below.

The reviewer also raises the concern that, regardless of practical implications, our results may amount to an incremental advance over those of (HS15) by way of similar techniques. We disagree with this assessment on both counts: our main result (Thm. 1) goes far beyond a straightforward extension of that in (HS15) to the noisy case and this required a significant deviation from their approach.

It is understandable that the reviewer may have thought otherwise, considering how the proof of Thm. 1 is presented in the manuscript. As observed by our second reviewer, however, the extension to the noisy case did indeed require a novel combination of several results in the literature. Specifically, the main difficulty was to generalize Lemma 1 to the case where the k-dimensional subspaces spanned by corresponding (through the map /pi) submatrices of A and B are assumed only to be proximal (small ?d?), and not identical as in (HS15). In contrast to the noiseless case, here it must be explicitly demonstrated that this

proximity relation is propagated through the repeated intersections of these submatrix spans all the way down to the spans of dictionary elements themselves. We designed and proved Lemma 3 to address this issue, which draws its bound from the convergence guarantees of an alternating projections algorithm first proposed by von Neumann. This result, combined with a little known fact about the distance ?d? requiring proof in (M10), constitute the more obscure components of the deduction in Eq. (26). To reiterate, this step is completely trivial in the noiseless case and required no mention for the inductive steps taken in (HS15).

Our proof of Lemma 1 diverges perhaps even more significantly from the approach taken in (HS15) by way of Lemma 4. Key to our reduction of the sample complexity given in (HS15) by an exponential factor is the introduction of a combinatorial design (the "singleton intersection property") for support sets. Since in this case the map /pi from supports in the hypergraph to ([m] /choose k) is not surjective, one can not apply the same inductive method as in (HS15), which freely chooses supports in the codomain to intersect at (k-1) nodes and map back to some corresponding (k-1)-sized intersection of supports in the domain. Instead, we demonstrate the surprising fact that by pigeonholing the images of supports in the (SIP-satisfying) hypergraph H, one can still guarantee a bijection between the nodes and therefore the subspaces spanned by individual dictionary elements. We note that it was necessary to forgo inductive methods altogether in order to prove this fact for all hypergraphs satisfying the SIP; otherwise, we would require that the supports in the hypergraph have intersections of size k? for every k? ¡ k (e.g., this is not the case for the small SIP example we give consisting of the rows and columns of nodes arranged in a square grid). Again, to our surprise, it so happens that this new induction-less argument easily generalizes to the case where B has an arbitrary number of columns, in which case we find that a one-to-one correspondence exists between a subset of columns of A and B of a size that has a nice closed-form expression for regular SIP hypergraphs.

To be clear, the new perspective we take on the problem yields the following powerful conclusions beyond those of a straightforward extension of (HS15) to the noisy case:

1) An extension to the case where the number of dictionary elements is unknown: The results of (HS15) only apply to the case where the matrix B has the same number of columns as A. We forgo this assumption and show that B must have at least as many columns as A and contains (up to noise) a subset of the columns of A. The size of this subset depends on a simple relation between the number of columns of B and the regularity of the support-set hypergraph.

2) A significant reduction in sample complexity: To identify the n x m generating dictionary A, (HS15) require that data be sampled from every k-dimensional subspace spanned by the m columns of A (that is, (m /choose k) subspaces in total). We show that the data need only be sampled from m subspaces in total (e.g. those supported by consecutive intervals of length k in some cyclic order on [m]) and in some cases as few as 2/sqrt(m) (e.g. when m = k**2, take the supports to be the rows and columns of [m] arranged in a square grid). Moreover, if the size of the support set of reconstructing codes is known to be polynomial in m and k, then the pigeonholing argument in the proof of Thm. 1 requires only a polynomial number of samples distributed over a polynomial number of supports; thus, N is then poly-

nomial in /bar m, k. This point was only hinted at in the Discussion section of our original submission, but we have included it in the updated manuscript to make clear the power of our approach over that taken in (HS15).

3) No spark condition requirement for the generating matrix A: One of the mathematically significant contributions made by (HS15) was to forgo the constraint that the recovery matrix B also satisfy the spark condition, in contrast to all previously known uniqueness results which (either explicitly or implicitly) made this assumption. Our proof is powerful enough to show that, in fact, even the matrix A need not satisfy the spark condition to be identifiable! Rather, it need only be injective on the union of subspaces with supports that form sets in a hypergraph satisfying the singleton intersection property (SIP). (For example, consider the matrix A = [e-1,.., e-5, v] where v = e-1 + e-3 + e-5, and take H to be the set of all consecutive pairs of [m] arranged in cyclic order. Then A satisfies the assumptions of Thm. 1 for dictionary recovery without satisfying the spark condition.) We had omitted this point in our original submission of the manuscript to keep things simple, but we have decided to include this fact in our revision. This also required us to redefine the restricted matrix lower bound L-k to be in terms of a hypergraph H (L-H in the revised manuscript), which is an interesting object for further study in its own right.

We must also reiterate here that our solution to Prob. 2, that of most interest to practitioners of dictionary learning methods, is to our knowledge the first of its kind in both the noise-free and noisy domains. We recognize that this was not clearly communicated in the Discussion section of our submitted manuscript.

Finally, we should mention that our main mathematical results justify the neurally plausible model of bottleneck communication between brain regions, first explained in depth in this NIPS paper (IHS10) and then in a review of sparse linear coding applications to neuroscience (GS12).

REVIEWER:The paper could also compare the sample complexity results on N with those given in the noise-free case. For example, Table I in reference (18) gives a comparison of sample complexity requirements for different conditions in the noise-free case. Where would the results of this paper stand in that table?

RESPONSE: In brief, since all of the theorems and corollaries in (HS15) are corollaries of our theorems, all sample complexities improve by our work. The main point of difference between our statements about sample complexity and those of (HS15) is the assumed constraint on the underlying support set of sparse codes x-i. Our theory of combinatorial designs (the ?singleton intersection property?) requires only that there be a sufficient number of x-i drawn from each support in a hypergraph satisfying the SIP, whereas the theory in (HS15) requires that a sufficient number of x-i be drawn from every support of size k in [m]. Below, we make an explicit comparison with Table I in (HS15) row by row, assuming for our results that sufficient data have been sampled from all supports in a hypergraph H satisfying the SIP:

ROW I. Our result here is —H—(k-1)(m /choose k) + 1. We improve on the result in (HS15) by an exponential factor since our theory does not require that data be sampled with

every possible support of size k in [m]; as noted in our manuscript, for every k ¡ m there exists a hypergraph H with —H— = m that satisfies the SIP.

ROW 2. Our result here is again —H—(k-1)(m /choose k) + 1 with certainty, not with almost certainty (i.e. probability 1) as in (HS15). Here, the authors in (HS15) have applied probabilistic arguments to achieve an almost certain result with a sample complexity on the same order as that for which we have achieved a certain result by means of our theory of combinatorial designs.

ROW 3. We cannot make a direct comparison here because we have not calculated the probability that a random set of supports satisfy the SIP (see our response to a similar question by the second reviewer). We think this is an interesting problem for the community to solve; regardless, the result of (HS15) here is still implied by our more general theory.

ROW 4. Our result here is —H—(k-1)(m /choose k) + 1 with probability 1. Here is the only case where the sample complexity stated in (HS15) is technically better than ours, but it differs in flavor: their x-i are assumed to be distributed as (k+1) samples per support for every support of size k in [m], whereas ours supposes that (k-1)(m /choose k) codes x-i are distributed over each support in some hypergraph H satisfying the SIP. Regardless, the (technically) better result of (HS15) is still implied by our more general theory in the case /varepsilon = 0 with the addition of their probabilistic argument.

ROW 5: Again, we cannot make a direct comparison here for the reason stated in ROW 3.

To sum up, while we appreciate this question, we feel that since our results improve in every case except for one noise-free technicality (ROW 4), and since the results of (HS15) are all entailed anyway by our more general theory, an explicit comparison such as that which we provided above would not be the best use of our limited space in the manuscript; although we are willing to add a discussion about this if necessary. Many thanks for your careful review of the mathematics. We have tried and continue to try to be as clear, concise, and correct as we possibly can to elevate this work into the top echelon of applied mathematical theory papers.

??

REVIEWER: An editor who doesn't want to get suckered is always looking at papers to find a very specific 'gadget' or 'gadgets' that mark the distinction between the submitted work and 'obvious', 'trivial' work.

I can't really tell if the hypergraph construct is such a gadget. It's really only handled in passing and only two examples are mentioned.

———————————-

RESPONSE:

In our previous response, we described in detail how the definition of a new matrix lower bound induced by a hypergraph enables the following advancements in our understanding of Prob. 1 beyond an extension of previous results to the noisy case:

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, 3) the spark condition is not a necessary property of recoverable dictionaries.

Given these insights, we believe the term "gadget" strongly downplays the power of this construct. Certainly, we don't see it as being only treated in passing, as it is incorporated into every one of our theorems. The examples we provide briefly demonstrate the gains to be made from this discovery, so as to entice the community into fully exploring the ramifications of an underlying theory of hypergraphs in practice.

??????

REVIEWER: So I'm at a standstill. I would need to be convinced that you have actual gadgets that go beyond what I would have come up with and that the identifiability problem is dramatically different than what an 'obvious' or 'easy-to-guess' solution might say, by showing me something very concrete that I can understand. The lack of any explicit implementation on a computer on a specific example doesn't help.

———————————-

RESPONSE:

It seems to us that assessing a solution as "obvious" after-the-fact – when it has already informed one's intuition – is a bit unfair. It is always easy to guess, easier to guess wrong, and hardest to prove. Still, if unintuitive results are what sell these days, we offer a nice surprise for everyone who reads our paper and realizes how strange it is that they have never come across a definition of the problem akin to Def. 1 anywhere in the literature on dictionary learning before.

Problems 1 and 2 have been studied for two decades now, and no one has pointed out the relation between them, nor the existence of the underlying hypergraph structure. Our solution to Prob. 2, that of most interest to practitioners of dictionary learning methods, is the first of its kind in both the noise-free and noisy domains. If these solutions are the "easy-to-guess", "obvious" solutions, then so be it; we cannot change geometry. Actually, we prefer results for which intuition can play its role in lending credence to the truth.

PRACTICAL IMPLICATIONS:

REVIEWER: When is it possible to determine (numerically) if enough data has been gathered to obtain an accurate estimate of the underlying sparsity dictionary? ————————

RESPONSE: A statement of this very general nature requires that we calculate the probability that all of the conditions of Thm. 1 are satisfied by random data. We have provided an "almost certain" guarantee of this flavor in Cor 2, wherein the sparse vector supports are constrained to form a hypergraph satisfying the SIP. It would be useful and interesting to calculate furthermore the probability that random supports form a hypergraph satisfying the SIP. We have opted not to include these calculations due in part to space constraints, but also because we believe that what sets our work apart from the vast majority of results in the field is their deterministic nature, e.g. they do not depend on any kind of assumption about the particular random distribution from which the sparse supports, coefficients, or dictionary entries are drawn.

———————— Is there any new intuition to be had regarding the necessary sampling of the union of subspaces from this analysis (with respect to the intuition obtained from existing uniqueness results)? ————————

RESPONSE: Yes, there is indeed. While all existing uniqueness results require a sufficient number of samples be drawn from each of (m /choose k) possible supports, we have shown that it is enough that a sufficient number of samples be drawn from every support in a subset of drastically smaller size, provided this subset forms a regular hypergraph satisfying the SIP. As was pointed out in our manuscript, a support set of size m satisfying these criteria can always be constructed for any k ¡ m (e.g. take the consecutive intervals of length k in some cyclic order on [m]), and in certain cases such a set can be constructed from as few as 2/sqrt(m) supports (e.g. when m = k**2, take the supports to be the rows and columns of [m] arranged in a square grid).

??

REVIEWER: This is at best about local stability, rather than stability. The size of the neighborhood where you're getting stability is presumably almost infinitesimal in general. While I agree that the paper is written in a quantitative way and has as you say constants which can be quantitative, it is I would say after a few readings, more seemingly-quantitative-but-actually-purely-qualitative. If it could be shown to be surprisingly effective in some specific case, I might withdraw from this position.

———————————-

RESPONSE:

[Files included: 2017-06210R-AuthorResponseManuscript.pdf, compute-C2.py, decomp-svd.py, sample-complexity.py, C2-Grid-m64-n32-k8-r2-nt500.pdf, C2-Grid-m256-n128-k16-r2-nt500.pdf, C2-Grid-m1024-n512-k32-r2-nt50.pdf, prob-vs-samples.png, samples-vs-m.png]

We are somewhat confused by this statement as we understand even the tightest possible noise bound to be necessarily "local" and approaching zero as the domain-restricted matrix lower-bound vanishes, e.g. as the number of dictionary elements grows. One need only consider the case k=1, where every dictionary element spans a ray through the origin, and each identifiable datum generated from the model of Eq. 1 lies within a cylinder of radius /eta around one (and only one) of these rays. Clearly, for bounded data in a finite dimensional space, this radius cannot remain finite in the limit of infinitely many distinct rays.

The issue seems to rather be that the bound (defined using the constant C1, which depends in turn on the constant C2) is, in general, so small that it is "practically infinitesimal". As is the case for any deterministic guarantee, we must entertain the "worst-case" scenario, where noise is isotropic and "what can go wrong, will go wrong?. Consequently, some degree of pessimism is forced upon us here, whereas calculations specific to the data at hand would consider the probability distribution of confounding noise and likely yield more forgiving probabilistic bounds. To respond to this valid concern nonetheless, we have investigated our deterministic constants more thoroughly.

The denominator of C1 involves a relatively standard quantity in the field of compressed sensing (the "restricted isometry constant"), and it is known to be reasonable for many random distributions generating original dictionaries A and codes. The constant C2, on the other hand, depends on a much less well studied quantity /xi computed from the "Friedrichs angle" between certain spans of A's columns.

Upon re-examination of our proof, we have determined that our expression for C2 was egregiously sub-optimal. As it turns out, the minimum in the denominator of C2 need only be computed over subsets of H of size r+1 (where r is the regularity of the hypergraph) – not over all subsets of H. Moreover, the 2**—H— in the numerator – which in fact need only have been —H— in our previous submission – can actually be set to r+1 (e.g. for the SIP hypergraph consisting of consecutive intervals of length k in [m] we have —H— = m, whereas r=k).

To determine the practical utility of this adjusted constant, we have performed computer experiments (Python code "compute-C2.py", "decomp-svd.py" included) calculating it for some generic matrices with example hypergraphs from our paper's introduction (i.e. the rows and columns formed by arranging [m] into a square grid). The results suggest that the bound is actually quite reasonable; certainly, it is not ?practically infinitesimal" (Figures "C2-Grid-m64-n32-k8-r2-nt500.pdf", "C2-Grid-m256-n128-k16-r2-nt500.pdf", "C2-Grid-m1024-n512-k32-r2-nt50.pdf", included over different latent dimensions m=64,256,1024).

REVIEWER: I can't really tell if your cardinality bound on the number of samples needed for a stable representation is one such gadget. I am unable on my own to conjure up an example where I might have thought an exponential number of samples would be required but you show me very explicitly that no, a dramatically smaller number is required.

RESPONSE:

If there is any one "gadget" to which we may credit our results, it is the pigeonhole principle, which we have applied in a way (see Lem. 4) that demonstrates that only a polynomial number of sparse supports are necessary in general for stable identification of the generating dictionary. In our view, this lends much more legitimacy to the use of the sparse linear coding model in practice, where data in general are unlikely (if ever) to exhibit the exponentially many possible k-wise combinations of dictionary elements, as all previous results have required.

It may very well be impossible to to exorcise exponentiality from the number of required samples in the deterministic or almost-certain case. However, our guarantees can easily be extended to hold with some probability for any number of samples by appealing instead to a probabilistic pigeonholing at the point in our proof of Thm. 1 where the deterministic pigeonhole principle is applied to demonstrate that for every S in H, there exist k vectors x-i supported on S whose corresponding /bar x-i all share the same support. (A famous example of such an argument is the counter-intuitive "birthday paradox", which demonstrates that the probability of two people having the same birthday in a room of twenty-three is in fact greater than 50pcnt.) This point in the proof then has some probability of success, which must occur for all S in H in order for the proof as a whole to be valid.

In this spirit, we have computed for all sample sizes up to our deterministic sample complexity the probability with which our guarantees still hold when H is one of the two example hypergraphs from our paper's introduction, the set of consecutive intervals of length k in a cyclic order on [m] (Figure "prob-vs-samples.png", Python code "sample-complexity.py").

The probability of our guarantees holding saturates when the number of samples reaches only a fraction of the deterministic sample complexity.

We have also computed the number of samples required for our guarantees to hold with probability 99.9percent for fixed k as m increases (Figure "samples-vs-m.png"). The ratio of the number of samples required for this 99.9percent guarantee with respect to the number of samples required by the deterministic 100percent guarantee of Thm. 1 tends to zero as m increases.

It would be a simple matter to add discussion of these facts to the manuscript, if so desired.

REVIEWER: In my opinion, while the extension from exact to noisy stability of dictionary learning (DL) is significant, the fact that the analysis relies on metrics of the data that are not feasible to compute limits its impact to the scientific community beyond computer science and applied mathematics. While the authors state in their response that their results validate the extensive successful use of DL in practice, there seems to be little impact to this given that the methodology is already in widespread use; instead, practical criteria that allows practitioners to establish whether the data model obtained from DL is optimal or not would have very high impact. My questions in the review were probing whether any contribution of this type was present, and the responses appear to point toward a negative answer.

RESPONSE:

What sets our work apart from the vast majority of results in the field is that they are deterministic, and do not depend on any kind of assumption about the particular random distribution from which the sparse supports, coefficients, or dictionary entries are drawn. Consequently, our paper directly justifies only "in principle" the inferences of those who apply dictionary learning methods to inverse problems in their research. But this is unavoidably the case for NP-hard problems.

In the realm of practicality, we have spelled-out the problem (i.e. estimate C1) for statisticians, who will derive from our deterministic guarantees the statistical criteria for inference in more domain-specific probabilistic models, and we have cut in half the work it takes a computer scientist to prove the consistency of any dictionary learning algorithm (i.e. prove that the algorithm converges to any solution encoding the data to within the epsilon in Eq. 8). Our work is the assist to these many impactful results to come, and (just as in hockey) we feel this deserves as much acknowledgement as any contingent goal.

THEORETICAL IMPLICATIONS

REVIEWER: Point (a) is important because when you say these results somehow validate various neuroscience ideas, I am saying in the back of my mind that you're bravely talking yourself on a ledge. Namely, nothing in neuroscience can really depend on infinitesimal stability. Trying to make a claim like this seems to undermine your credibility with me. It seems the referees had the same problem.

RESPONSE:

Our response to point (a) aside, and though we are not inclined to assume that neural circuits are incapable of extreme precision when necessary, we actually share your skepticism here regarding the modern theory of "sparse coding" (O04) for efficient representation of sensory input in brains (vision: (O96), (H96), (B97), (vH98); audio: (B96), (S06), (C12)), as well as corresponding models of bottleneck communication between neurons that utilize ideas from compressed sensing ((C10), (I10), (G12)). Nonetheless, our paper confirms the well-posedness of the central noisy sensory coding problem that specialists in these fields have suggested neural circuits might be solving, and we have verified (for the first time) that the bottleneck communication model represents a neurally plausible way of faithfully transmitting sparse sensory representations through a noisy bandwidth-limited channel.

If the editors are still reluctant to become complicit in propagating unproven hypotheses of neural computation, we have no qualms downplaying this application in the manuscript. Perhaps instead, we could elaborate more on implications for the repeatability of discoveries in experimental science contingent on machine learning, which includes an explanation for the universality of the above results (e.g. "Gabors") in sensory representation (independent of any particular theory of brain computation). For example, over the years there have been many appeals to dictionary learning with sparseness constraint for uncovering latent structure in data (e.g. forgery detection (H10), (O10); brain recordings (J01), (A14), (L16); gene expression (Wu16)), several of which appear in PNAS.

REVIEWER: It's not clear why exact uniqueness should be so important for neuroscience. It seems more likely that mere similarity is what's maybe important. In the same sense it's not clear why formal Lipschitz stability in the mathematical sense should be so important for neuroscience. It seems more important that some sort of qualitative similarity should persist under perturbations.

RESPONSE:

Unfortunately, how brains work is still largely a mystery. There are, however, some ideas for why certain models are more appealing than others. Here is how we rationalize the application of unsupervised sparse coding to the theory of bottleneck communication:

Suppose some quantities of interest are encoded in the sparse activity of neurons in a sending region (it has been proposed that neural activity in certain brain regions is "sparse", e.g. for energy efficiency). These quantities are to be transmitted to some other region through as few wires (axons) as possible, e.g. due to space constraints inside the skull.

The hypothesis is that the brain, up-to date on the latest fad, solves this problem by applying a (noisy) "random" projection into the lower-dimensional space spanned by these axons. The neurons in the receiving region are then tasked with decoding the quantities of interest from this compressed signal. Suppose, however, that they cannot read out this information directly from the compressed signal. Rather, the neurons in the receiving region must first reproduce to some extent the original sparse pattern of activity before they can decode from it the quantities of interest (perhaps a sparse representation is necessary or

advantageous in the receiving region, just as in the sending region). Moreover - and this is central to the hypothesis - they must accomplish this feat without knowledge of the random projection applied by the sender, i.e. via dictionary learning with a sparseness constraint.

In this way, the persistence of a "qualitative similarity", modeled mathematically as the formal similarity between the decodable quantities of interest in the two regions, is contingent on the uniqueness and stability of sparse representations. We have proven that any dictionary learning procedure implemented in the receiving region (biologically plausible algorithms are an active area of research (P15)) that does "well enough" will indeed yield a sparse activity pattern that is similar (at least, up to inherent relabelling ambiguity) to the original pattern in the sending region, as required.

It is entirely possible, and in our opinion very likely, that the encoding and decoding procedures implemented in nervous tissue are far more sophisticated. The science is just not there yet. At the very least, we have demonstrated that the only published hypothesis regarding how this could be done is well-founded, mathematically.

REVIEWER: The perturbation bound is used by the authors to provide some explanation for why, in practice, different methods lead to similar dictionaries. Can that also be used to explain the second (bold) part of their observation that ?[these waveforms] appear in dictionaries learned by a variety of algorithms trained with respect to different natural image datasets.?? Just curious.

RESPONSE: To be clear, this (bold) observation is just an observation which motivates the work. Imagining natural image patches to form a population which satisfies (1) with respect to some ground truth dictionary A, we have shown that there exists a sufficient number of samples to uniquely identify it or something close to it via approximate solutions. Given this fundamental stability and uniqueness of the problem solutions, the fact that a variety of methods geared to approximately solve Problems 1 and 2 all seem to capture similar structure is much less surprising.

FUTURE DIRECTIONS

NOTE: Structured dictionaries? Just derive the bounds specific to them. Could be efficiently computable in specific cases.

REVIEWER

As l0 norm is intractable, does all the properties still hold for l1 norm minimization? Can we solve it efficiently (polynomial time) to global optimality? Similar issue for other works, such as https://arxiv.org/pdf/1807.05595.pdf

RESPONSE: We make no claims in this paper about l1-norm minimization, though we are hopeful a clever reader may find some way to build off our results and derive related guarantees for this continuous relaxation of the sparse coding problem.

TODO : can we deterministically generate matrices injective on SIP hyper graphs? The (mCk)-hypergraph is one such possibility, so maybe not in general? But for special cases (convolution matrices..?)

0.7 Appendix

We prove Lem. 1 after the following auxiliary lemmas.

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

Proof. By induction, it is enough to prove the case $\ell = 2$. Clearly, for any map f, if $w \in f(U \cap V)$ then $w \in f(U)$ and $w \in f(V)$; hence, $w \in f(U) \cap f(V)$. If $w \in f(U) \cap f(V)$, then $w \in f(U)$ and $w \in f(V)$; thus, $w = f(u) = f(v)$ for some $u \in U$ and $v \in V$, implying $u = v$ by injectivity of f. It follows that $u \in U \cap V$ and $w \in f(U \cap V)$. □

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

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

$$dist(\mathbf{u}, V) \leq \frac{1}{1 - \xi(\mathcal{V})} \sum_{i=1}^{k} dist(\mathbf{u}, V_i). \tag{22}$$

Proof. Recall the projection onto the subspace $V \subseteq \mathbb{R}^m$ is the mapping $\Pi_V : \mathbb{R}^m \to V$ that associates with each $\mathbf{u}$ its unique nearest point in V; i.e., $\|\mathbf{u} - \Pi_V \mathbf{u}\|_2 = \text{dist}(\mathbf{u}, V)$. By repeatedly applying the triangle inequality, we have:

$$\begin{aligned} \|\mathbf{u} - \Pi_V \mathbf{u}\|_2 &\leq \|\mathbf{u} - \Pi_{V_k}\mathbf{u}\|_2 + \|\Pi_{V_k}\mathbf{u} - \Pi_{V_k}\Pi_{V_{k-1}}\mathbf{u}\|_2 \\ &\quad + \cdots + \|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\mathbf{u} - \Pi_V\mathbf{u}\|_2 \\ &\leq \sum_{\ell=1}^{k} \|\mathbf{u} - \Pi_{V_\ell}\mathbf{u}\|_2 + \|(\Pi_{V_k}\cdots\Pi_{V_1} - \Pi_V)\mathbf{u}\|_2, \end{aligned} \tag{23}$$

where we have also used that the spectral norm of the orthogonal projections Π_{V_ℓ} satisfies $\|\Pi_{V_\ell}\|_2 \leq 1$ for all ℓ.

It remains to bound the second term in (23) by $\xi(\mathcal{V})\|\mathbf{u} - \Pi_V\mathbf{u}\|_2$. First, note that $\Pi_{V_\ell}\Pi_V = \Pi_V$ and $\Pi_V^2 = \Pi_V$, so we have $\|(\Pi_{V_k}\cdots\Pi_{V_1} - \Pi_V)\mathbf{u}\|_2 = \|(\Pi_{V_k}\cdots\Pi_{V_1} - \Pi_V)(\mathbf{u} - \Pi_V\mathbf{u})\|_2$. Consequently, inequality (22) follows from [6, Thm. 9.33]:

$$\|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\mathbf{x} - \Pi_V\mathbf{x}\|_2 \leq z\|\mathbf{x}\|_2, \quad \text{for all } \mathbf{x}, \tag{24}$$

with $z^2 = 1 - \prod_{\ell=1}^{k-1}(1 - z_\ell^2)$ and $z_\ell = \cos\theta\left(V_\ell, \cap_{s=\ell+1}^{k} V_s\right)$ (recall θ from Def. 4), after substituting $\xi(\mathcal{V})$ for z and rearranging terms. □

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

$$|\cap \pi(\mathcal{G})| \leq |\cap \mathcal{G}|, \quad \textit{for } \mathcal{G} \in \binom{\mathcal{H}}{r} \cup \binom{\mathcal{H}}{r+1}, \tag{25}$$

then $\overline{m} \geq m$; and if $\overline{m}(r-1) < mr$, the map $i \mapsto \cap_{S\in\sigma(i)}\pi(S)$ is an injective function to $[\overline{m}]$ from some $J \subseteq [m]$ of size $m - (r-1)(\overline{m}-m)$ (recall σ from Def. 2).

Proof. Consider the following set: $T_1 := \{(i, S) : i \in \pi(S), S \in \mathcal{H}\}$, which numbers $|T_1| = \sum_{S\in\mathcal{H}} |\pi(S)| \geq \sum_{S\in\mathcal{H}} |S| = \sum_{i\in[m]} \deg_{\mathcal{H}}(i) = mr$ by r-regularity of $\mathcal{H}$. Note that $|T_1| \leq \overline{m}r$; otherwise, pigeonholing the tuples of T_1 with respect to their $\overline{m}$ possible first elements would imply that more than r of the tuples in T_1 share the same first element. This cannot be the case, however, since then some $\mathcal{G} \in \binom{\mathcal{H}}{r+1}$ formed from any $r+1$ of their second elements would satisfy $\cap\pi(\mathcal{G}) \neq 0$; hence, $|\cap\mathcal{G}| \neq 0$ by (25), contradicting r-regularity of $\mathcal{H}$. It follows that $\overline{m} \geq m$.

Suppose now that $\overline{m}(r-1) < mr$, so that $p := mr - \overline{m}(r-1)$ is positive and $|T_1| \geq \overline{m}(r-1) + p$. Pigeonholing T_1 into $[\overline{m}]$ again, there are at least r tuples in T_1 sharing some first element; that is, for some $\mathcal{G}_1 \subseteq \mathcal{H}$ of size $|\mathcal{G}_1| \geq r$, we have $|\cap \pi(\mathcal{G}_1)| \geq 1$ and (by (25)) $|\cap \mathcal{G}_1| \geq 1$. Since no more than r tuples of T_1 can share the same first element, we in fact have $|\mathcal{G}_1| = r$. It follows by r-regularity that $\mathcal{G}_1$ is a star of $\mathcal{H}$; hence, $|\cap\mathcal{G}_1| = 1$ by the SIP and $|\cap\pi(\mathcal{G}_1)| = 1$ by (25).

If $p = 1$, then we are done. Otherwise, define $T_2 := T_1 \setminus \{(i, S) \in T_1 : i = \cap\pi(\mathcal{G}_1)\}$, which contains $|T_2| = |T_1| - r \geq (\overline{m}-1)(r-1) + (p-1)$ ordered pairs having $\overline{m}-1$ distinct first indices. Pigeonholing T_2 into $[\overline{m}-1]$ and repeating the above arguments produces the star $\mathcal{G}_2 \in \binom{\mathcal{H}}{r}$ with intersection $\cap\mathcal{G}_2$ necessarily distinct (by r-regularity) from $\cap\mathcal{G}_1$. Iterating this procedure p times in total yields the stars $\mathcal{G}_i$ for which $\cap\mathcal{G}_i \mapsto \cap\pi(\mathcal{G}_i)$ defines an injective map to $[\overline{m}]$ from $J = \{\cap\mathcal{G}_1, \ldots, \cap\mathcal{G}_p\} \subseteq [m]$. □

Proof of Lem. 1. We begin by showing that $\dim(\boldsymbol{\mathcal{B}}_{\pi(S)}) = \dim(\boldsymbol{\mathcal{A}}_S)$ for all $S \in \mathcal{H}$. Note that since $\|\mathbf{Ax}\|_2 \leq \max_j \|\mathbf{A}_j\|_2 \|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_1 \leq \sqrt{k}\|\mathbf{x}\|_2$ for all k-sparse $\mathbf{x}$, by (5) we have $L_2(\mathbf{A}) \leq \max_j \|\mathbf{A}_j\|_2$ and therefore (as $0 \leq \xi < 1$) the right-hand side of (15) is less than one. From (13), we have $|\pi(S)| \geq \dim(\boldsymbol{\mathcal{B}}_{\pi(S)}) \geq \dim(\boldsymbol{\mathcal{A}}_S) = |S|$, the final equality holding by injectivity of $\mathbf{A}_S$. As $|\pi(S)| = |S|$, the claim follows. Note, therefore, that $\mathbf{B}_{\pi(S)}$ has full-column rank for all $S \in \mathcal{H}$.

We next demonstrate that (25) holds. Fixing $\mathcal{G} \in \binom{\mathcal{H}}{r} \cup \binom{\mathcal{H}}{r+1}$, it suffices to show that $d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) < 1$, since by (13) we then have $|\cap\pi(\mathcal{G})| = \dim(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}) \leq \dim(\boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) = |\cap\mathcal{G}|$, with equalities from the full column-ranks of $\mathbf{A}_S$ and $\mathbf{B}_{\pi(S)}$ for all $S \in \mathcal{H}$.[17] Observe that $d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) \leq d\left(\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})}, \cap\boldsymbol{\mathcal{A}}_{\mathcal{G}}\right)$ by (12), since trivially $\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})} \subseteq \cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})}$ and also

[17] Note that if ever $\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})} \neq /bfseries0$ while $\cap\mathcal{G} = \emptyset$, we would have $d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \mathbf{0}) = 1$. However, that leads to a contradiction.

$\boldsymbol{\mathcal{A}}_{\cap\mathcal{G}} = \cap\boldsymbol{\mathcal{A}}_{\mathcal{G}}$ by Lem. 2. Recalling Def. 3 and applying Lem. 3 yields:

$$\begin{aligned} d\left(\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})}, \cap\boldsymbol{\mathcal{A}}_{\mathcal{G}}\right) &\leq \max_{\mathbf{u}\in\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})},\ \|\mathbf{u}\|_2\leq 1} \sum_{S\in\mathcal{G}} \frac{\operatorname{dist}\left(\mathbf{u}, \boldsymbol{\mathcal{A}}_S\right)}{1-\xi(\boldsymbol{\mathcal{A}}_{\mathcal{G}})} \\ &= \sum_{S\in\mathcal{G}} \frac{d\left(\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_S\right)}{1-\xi(\boldsymbol{\mathcal{A}}_{\mathcal{G}})}, \end{aligned}$$

passing the maximum through the sum. Since $\cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})} \subseteq \boldsymbol{\mathcal{B}}_{\pi(S)}$ for all $S \in \mathcal{G}$, by (12) the numerator of each term in the sum above is bounded by $d\left(\boldsymbol{\mathcal{B}}_{\pi(S)}, \boldsymbol{\mathcal{A}}_S\right) = d\left(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\pi(S)}\right) \leq \varepsilon$, with the equality from (14) since $\dim(\boldsymbol{\mathcal{B}}_{\pi(S)}) = \dim(\boldsymbol{\mathcal{A}}_S)$. Thus, altogether:

$$d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) \leq \frac{|\mathcal{G}|\varepsilon}{1-\xi(\boldsymbol{\mathcal{A}}_{\mathcal{G}})} \leq \frac{C_2\varepsilon}{\max_j \|\mathbf{A}_j\|_2}, \tag{26}$$

recalling the definition of C_2 in (18). Lastly, since $C_2\varepsilon < L_2(\mathbf{A}) \leq \max_j \|\mathbf{A}_j\|_2$, we have $d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) \leq 1$ and therefore (25) holds.

Applying Lem. 4, the association $i \mapsto \cap_{S\in\sigma(i)}\pi(S)$ is an injective map $\overline{\pi} : J \to [\overline{m}]$ for some $J \subseteq [m]$ of size $m-(r-1)(\overline{m}-m)$, and $\mathbf{B}_{\overline{\pi}(i)} \neq \mathbf{0}$ for all $i \in J$ since the columns of $\mathbf{B}_{\pi(S)}$ are linearly independent for all $S \in \mathcal{H}$. Letting $\overline{\varepsilon} := C_2\varepsilon/\max_i \|\mathbf{A}_i\|_2$, it follows from (14) and (26) that $d\left(\boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{B}}_{\overline{\pi}(i)}\right) = d\left(\boldsymbol{\mathcal{B}}_{\overline{\pi}(i)}, \boldsymbol{\mathcal{A}}_i\right) \leq \overline{\varepsilon}$ for all $i \in J$. Setting $c_i := \|\mathbf{A}_i\|_2^{-1}$ so that $\|c_i\mathbf{A}\mathbf{e}_i\|_2 = 1$, by Def. 3 for all $i \in J$:

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

for $\overline{\varepsilon} < L_2(\mathbf{A}) \min_{i\in[m]} |c_i|$. But this is exactly the supposition in (10), with J and $\overline{\varepsilon}$ in place of $[m]$ and ε, respectively. The same arguments of the case $k = 1$ in Sec. 0.3 can then be made to show that for any $\overline{m} \times \overline{m}$ permutation and invertible diagonal matrices $\mathbf{P}$ and $\mathbf{D}$ with, respectively, columns $\mathbf{e}_{\pi(i)}$ and $\frac{\overline{c}_i}{c_i}\mathbf{e}_i$ for $i \in J$ (otherwise, $\mathbf{e}_i$ for $i \in [\overline{m}] \setminus J$), we have $\|\mathbf{A}_i - (\mathbf{BPD})_i\|_2 \leq \overline{\varepsilon}/|c_i| \leq C_2\varepsilon$ for all $i \in J$. □

Bibliography

[1] G. Agarwal et al. "Spatially distributed local fields in the hippocampus encode rat position". In: *Science* 344.6184 (2014), pp. 626–630.

[2] M. Aharon, M. Elad, and A. Bruckstein. "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them". In: *Linear Algebra Appl.* 416.1 (2006), pp. 48–67.

[3] E. Arias-Castro, E. Candes, and M. Davenport. "On the fundamental limits of adaptive sensing". In: *IEEE Trans. Inf. Theory* 59.1 (2013), pp. 472–481.

[4] A. Bell and T. Sejnowski. "The "independent components" of natural scenes are edge filters". In: *Vision Res.* 37.23 (1997), pp. 3327–3338.

[5] J. Blanchard, C. Cartis, and J. Tanner. "Compressed sensing: How sharp is the restricted isometry property?" In: *SIAM Rev.* 53.1 (2011), pp. 105–125.

[6] F. Deutsch. *Best approximation in inner product spaces*. Springer Science & Business Media, 2012.

[7] D. Donoho and A. Flesia. "Can recent innovations in harmonic analysis 'explain' key findings in natural image statistics?" In: *Network Comp. Neural* 12.3 (2001), pp. 371–393.

[8] David L Donoho and Michael Elad. "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization". In: *Proceedings of the National Academy of Sciences* 100.5 (2003), pp. 2197–2202.

[9] M. Duarte et al. "Single-Pixel Imaging via Compressive Sampling". In: *IEEE Signal Process. Mag.* 25.2 (2008), pp. 83–91.

[10] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge university press, 2012.

[11] G. Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.

[12] Surya Ganguli and Haim Sompolinsky. "Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis". In: *Annu. Rev. Neurosci.* 35 (2012), pp. 485–508.

[13] L. Gao et al. "Single-shot compressed ultrafast photography at one hundred billion frames per second". In: *Nature* 516.7529 (2014), pp. 74–77.

[14] P. Georgiev, F. Theis, and A. Cichocki. "Sparse component analysis and blind source separation of underdetermined mixtures". In: *IEEE Trans. Neural Netw.* 16 (2005), pp. 992–996.

[15] I Goodfellow, J Shlens, and C Szegedy. "Explaining and harnessing adversarial examples". In: *Proc. International Conference on Learning Representations, 11 pp.* (2014).

[16] J. Grcar. "A matrix lower bound". In: *Linear Algebra Appl.* 433.1 (2010), pp. 203–220.

[17] Jacques Hadamard. "Sur les problèmes aux dérivées partielles et leur signification physique". In: *Princeton University Bulletin* 13.49-52 (1902), p. 28.

[18] J. van Hateren and A. van der Schaaf. "Independent component filters of natural images compared with simple cells in primary visual cortex". In: *Proc. R Soc. Lond. [Biol.]* 265.1394 (1998), pp. 359–366.

[19] C Hillar and F Sommer. "When can dictionary learning uniquely recover sparse data from subsamples?" In: *IEEE Trans. Inf. Theory* 61.11 (2015), pp. 6290–6297.

[20] J. Hughes, D. Graham, and D. Rockmore. "Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder". In: *Proc. Natl. Acad. Sci.* 107.4 (2010), pp. 1279–1283.

[21] J. Hurri et al. "Image feature extraction using independent component analysis". In: *Proc. NORSIG '96 (Nordic Signal Proc. Symposium).* 1996, pp. 475–478.

[22] Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications". In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[23] 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.* 2010, pp. 910–918.

[24] T.-P. Jung et al. "Imaging brain dynamics using independent component analysis". In: *Proc. IEEE* 89.7 (2001), pp. 1107–1122.

[25] Tosio Kato. *Perturbation theory for linear operators.* Vol. 132. Springer Science & Business Media, 2013.

[26] Chen Kong and Simon Lucey. "Prior-less compressible structure from motion". In: *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2016, pp. 4123–4131.

[27] Y.-B. Lee et al. "Sparse SPM: Group Sparse-dictionary learning in SPM framework for resting-state functional connectivity MRI analysis". In: *Neuroimage* 125 (2016), pp. 1032–1045.

[28] Y. Li, A. Cichocki, and S.-I. Amari. "Analysis of sparse representation and blind source separation". In: *Neural Comput.* 16.6 (2004), pp. 1193–1234.

[29] M. Lustig et al. "Compressed sensing MRI". In: *IEEE Signal Process. Mag.* 25.2 (2008), pp. 72–82.

[30] I. Morris. "A rapidly-converging lower bound for the joint spectral radius via multiplicative ergodic theory". In: *Adv. Math.* 225.6 (2010), pp. 3425–3445.

[31] B. Olshausen and D. Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583 (1996), pp. 607–609.

[32] Bruno A Olshausen and Michael R DeWeese. "Applied mathematics: The statistics of style". In: *Nature* 463.7284 (2010), p. 1027.

[33] C. Pehlevan and D. Chklovskii. "A normative theory of adaptive dimensionality reduction in neural networks". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2269–2277.

[34] M. Rehn and F. Sommer. "A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields". In: *J. Comput. Neurosci.* 22.2 (2007), pp. 135–146.

[35] C. Rozell et al. "Neurally plausible sparse coding via thresholding and local competition". In: *Neural Comput.* 20.10 (2008), pp. 2526–2563.

[36] J Sun, Q Qu, and J Wright. "Complete dictionary recovery over the sphere I: Overview and the geometric picture". In: *IEEE Trans. Inf. Theory* (2016), pp. 853 –884.

[37] A. Tillmann. "On the Computational Intractability of Exact and Approximate Dictionary Learning". In: *IEEE Signal Process. Lett.* 22.1 (2015), pp. 45–49.

[38] A. Tillmann and M. Pfetsch. "The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing". In: *IEEE Trans. Inf. Theory* 60.2 (2014), pp. 1248–1259.

[39] Rene Vidal, Yi Ma, and Shankar Sastry. "Generalized principal component analysis (GPCA)". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27.12 (2005), pp. 1945–1959.

[40] Z. Wang et al. *Sparse coding and its applications in computer vision*. World Scientific, 2015.

[41] S. Wu et al. "Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks". In: *Proc. Natl. Acad. Sci.* 113.16 (2016), pp. 4290–4295.

[42] Zheng Zhang et al. "A survey of sparse representation: algorithms and applications". In: *Access, IEEE* 3 (2015), pp. 490–530.