When is Sparse Coding Well-Posed?

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

***Abstract*—Sparse coding has exposed underlying structure in many kinds of natural data. However, given the multitude of algorithms implementing this strategy, claims of "true" latent discovery require the backing of universal theorems guaranteeing statistical consistency. Here, we prove that for almost all diverse enough datasets generated under this model, sparse coding identifies the original dictionary and codes up to an error commensurate with measurement noise. Applications are given to smoothed analysis, neuroscience, and engineering.**

***Index Terms*—Bilinear inverse problem, matrix factorization, identifiability, dictionary learning, sparse coding, compressed sensing, blind source separation, sparse component analysis**

I. Introduction

Ever since sparse coding of natural images reproduced response properties of neurons in mammalian primary visual cortex [1], learning sparse representations of vector-valued data has become a part of many signal processing and machine learning applications (see [2] for a comprehensive review). In the *sparse coding* (or *dictionary learning*) model, each point in a dataset $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\} \subset \mathbb{R}^n$ is approximated as a linear combination of at most k vectors drawn from a learned *dictionary* $\mathcal{A} \subset \mathbb{R}^n$, where $k < |\mathcal{A}| \ll N$.

Many algorithms have been designed to infer the parameters of this model, and it is tempting to interpret their output as approximating "ground truth" when it is thought to exist (e.g., [3]). It may be, however, that several qualitatively different solutions are in fact consistent with the data. The main finding of this work is that any dictionary satisfying the spark condition (2) from compressed sensing (CS) is uniquely identifiable from enough generic noisy sparse linear combinations of its elements up to an error linear in the noise (Thm. 1). In fact, provided n, m, and k satisfy the nearly-optimal CS inequality (6), then in almost all cases the dictionary learning problem is well-posed in the sense of Hadamard [4] (Cor. 2).

These algorithm-independent guarantees can also be extended to the case when only an upper bound on the size of $\mathcal{A}$ is known (Thm. 3). The explicit criteria under which these results hold can serve as theoretical tools in the analysis of sparse coding routines, some of which now provably converge to a global solution when it exists (see [5, Sec. I-E] for a brief discussion of the state-of-the-art in these algorithms).

We pose the sparse coding problem more precisely as follows. Fix a dictionary represented as the columns A_j of a matrix $A \in \mathbb{R}^{n \times m}$ and suppose Z consists of measurements:

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

for k-*sparse* $\mathbf{a}_i \in \mathbb{R}^m$ having at most k nonzero entries and *noise* $\mathbf{n}_i \in \mathbb{R}^n$ with ℓ_2-norm at most η. The noise represents our combined worst-case uncertainty in measuring $A\mathbf{a}_i$.

Problem 1 (Sparse Coding)**.** *Find $B \in \mathbb{R}^{n \times m}$ and k-sparse $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^m$ such that $|\mathbf{z}_i - B\mathbf{b}_i|_2 \le \eta$ for $i = 1, \ldots, N$.*

Note that any particular solution to this problem in fact represents a whole class of equivalent solutions BPD and $D^{-1}P^\top \mathbf{b}_i$, where $P \in \mathbb{R}^{m \times m}$ is any permutation matrix and $D \in \mathbb{R}^{m \times m}$ any invertible diagonal matrix. Since such a scaling and arbitrary ordering of dictionary elements represents a structurally equivalent model, it is natural to ask whether solutions to Problem 1 are unique up to this equivalence.

Previous work [6], [7], [8], [9] on the noiseless case $\eta = 0$ has shown that the solution (when it exists) is indeed unique in this sense provided the $\mathbf{a}_i$ are sufficiently diverse and the matrix A satisfies the *spark condition*:

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

which is evidently a necessary condition given that the $\mathbf{a}_i$ are known only to be k-sparse. Matrices of the form PD thus form the *ambiguity transformation group* inherent to the noiseless problem subject to these constraints [10].

We introduce the following terminology to handle $\eta > 0$.

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 $A \in \mathbb{R}^{n \times m}$ and k-sparse $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ are such that $\mathbf{y}_i = A\mathbf{a}_i$ for all i. This representation is **stable** if for every $\delta_1, \delta_2 \ge 0$, there exists $\varepsilon = \varepsilon(\delta_1, \delta_2) \ge 0$ (with $\varepsilon > 0$ when $\delta_1, \delta_2 > 0$) such that if a matrix $B \in \mathbb{R}^{n \times m}$ and k-sparse $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^m$ have $|A\mathbf{a}_i - B\mathbf{b}_i|_2 \le \varepsilon$ for all i, then there is a permutation matrix $P \in \mathbb{R}^{m \times m}$ and invertible diagonal matrix $D \in \mathbb{R}^{m \times m}$ such that for all $i = 1, \ldots, N$ and $j = 1, \ldots, m$:*

$$|A_j - (BPD)_j|_2 \le \delta_1 \quad \text{and} \quad |\mathbf{a}_i - D^{-1}P^\top \mathbf{b}_i|_1 \le \delta_2. \tag{3}$$

We ask here: *When does $Y \subset \mathbb{R}^n$ have a stable k-sparse representation in $\mathbb{R}^m$?* To see how a positive answer to this question informs the interpretation of solutions to Problem 1, suppose that $Y = \{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ has a stable k-sparse representation in $\mathbb{R}^m$, and fix δ_1, δ_2 to be the desired accuracy in recovery (3). Consider now any dataset Z generated as in (1) that has $\eta \le \frac{1}{2}\varepsilon(\delta_1, \delta_2)$. Then, any dictionary $B \in \mathbb{R}^{n \times m}$ and k-sparse $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^m$ solving Problem 1 approximate the original dictionary A and codes $\mathbf{a}_i$ (i.e., satisfy (3)).

In the next section, we give precise statements of our main results, which include an explicit form for $\varepsilon(\delta_1, \delta_2)$. We then prove our main theorem (Thm. 1) in Sec. III after listing some additional definitions and lemmas required for the proof, including our main tool from combinatorial matrix analysis

CG and CH supported, in part, by NSF grant IIS-1219212 and the Statistical and Applied Mathematical Sciences Institute, under NSF DMS-1127914.

(Lem. 1). Our proof is a refinement of the arguments in [9] to handle noise and to reduce the number of required samples from $N = k\binom{m}{k}^2$ to $N = m(k-1)\binom{m}{k} + m$. All other proofs are relegated to the supplement. Potential applications are discussed in the final section, Sec. IV.

II. Results

Before precisely stating our results, we explain how the spark condition (2) relates to the *lower bound* [11] of A, written $L(A)$, which is the largest number α such that $|A\mathbf{x}|_2 \geq \alpha|\mathbf{x}|_2$ for all $\mathbf{x} \in \mathbb{R}^m$. By compactness, every injective linear map has a nonzero lower bound; hence, if A satisfies (2), then every submatrix formed from $2k$ of its columns or less has a nonzero lower bound. We therefore define the following domain-restricted lower bound of A:

$$L_k(A) := \max\{\alpha : |A\mathbf{x}|_2 \geq \alpha|\mathbf{x}|_2 \text{ for all } k\text{-sparse } \mathbf{x} \in \mathbb{R}^m\}.$$

Clearly, $L_k(A) \geq L_{k'}(A)$ whenever $k < k'$, and for any A satisfying (2), we have $L_{k'}(A) > 0$ for all $k' \leq 2k$.

A *cyclic order* on $[m] := \{1, \ldots, m\}$ is an arrangement of $[m]$ in a circular necklace, and an *interval* in the order is any subset of contiguous elements. A vector $\mathbf{a} \in \mathbb{R}^m$ is said to be *supported* on $S \subseteq [m]$ when $\mathbf{a} \in \mathrm{Span}\{\{\mathbf{e}_i\}_{i\in S}\}$, where $\mathbf{e}_i$ are the standard basis vectors. Also, recall that M_j denotes the jth column of a matrix M. The following result gives a positive answer to our question from the introduction.

Theorem 1. *Fix n, m, and $k < m$. If $A \in \mathbb{R}^{n\times m}$ satisfies spark condition (2) and k-sparse $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ are such that for every interval of length k in some cyclic order on $[m]$ there are at least $(k-1)\binom{m}{k} + 1$ vectors $\mathbf{a}_i$ in general linear position (i.e., any k of them are linearly independent) supported there, then $Y = \{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ has a stable k-sparse representation in $\mathbb{R}^m$.*

Specifically, there exists a constant $C > 0$ for which the following holds for all $\varepsilon < \frac{L_2(A)}{\sqrt{2}}C^{-1}$. If any matrix $B \in \mathbb{R}^{n\times m}$ and k-sparse $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^m$ are such that $|A\mathbf{a}_i - B\mathbf{b}_i|_2 \leq \varepsilon$ for all $i \in [N]$, then for all $j \in [m]$:

$$|A_j - (BPD)_j|_2 \leq C\varepsilon, \tag{4}$$

for some permutation matrix $P \in \mathbb{R}^{m\times m}$ and invertible diagonal matrix $D \in \mathbb{R}^{m\times m}$. Moreover, if $\varepsilon < \varepsilon_0 := \frac{L_{2k}(A)}{\sqrt{2k}}C^{-1}$, then B also satisfies the spark condition and for all $i \in [N]$:

$$|\mathbf{a}_i - D^{-1}P^\top\mathbf{b}_i|_1 \leq \frac{\varepsilon}{\varepsilon_0 - \varepsilon}\left(C^{-1} + |\mathbf{a}_i|_1\right). \tag{5}$$

Remark 1. *Note that it was not assumed as given that B satisfy the spark condition. In fact, when $\varepsilon < \varepsilon_0$, we have $L_{2k}(BPD) \geq L_{2k}(A)\left(1 - \frac{\varepsilon}{\varepsilon_0}\right)$.*

As an important consequence, for sufficiently small reconstruction error, the original dictionary and codes are determined up to a commensurate error. Specifically, for $\delta_1, \delta_2 \geq 0$, Thm. 1 says that (3) is implied for any $\varepsilon < \varepsilon_0$ satisfying:

$$\varepsilon \leq \min\left(\delta_1 C^{-1}, \frac{\delta_2\varepsilon_0}{\delta_2 + C^{-1} + \max_{i\in[N]}|\mathbf{a}_i|_1}\right).$$

The constant C is explicitly defined in (7), below.

Corollary 1. *Given n, m, and $k < m$, there are $N = m(k-1)\binom{m}{k} + m$ vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ such that every matrix $A \in \mathbb{R}^{n\times m}$ satisfying (2) generates a set $Y = \{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ with a stable k-sparse representation in $\mathbb{R}^m$.*

It is straightforward to provide a probabilistic extension of Thm. 1 using the following fact in random matrix theory. The matrix $A \in \mathbb{R}^{n\times m}$ satisfies (2) with probability one provided:

$$n \geq \gamma k \log\left(\frac{m}{k}\right), \tag{6}$$

where γ is a positive constant dependent on the particular continuous distribution from which the entries of A are sampled i.i.d. (many ensembles suffice, e.g. [12, Sec. 4]).

In fact, the spark condition can be made explicit. Let A be the $n \times m$ matrix of nm indeterminates A_{ij}. When real numbers are substituted for all the A_{ij}, the resulting matrix satisfies (2) if and only if the following polynomial is nonzero:

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

where for any $S' \in \binom{[n]}{k}$ and $S \in \binom{[m]}{k}$, the symbol $A_{S',S}$ denotes the submatrix of entries A_{ij} with $(i,j) \in S' \times S$.

Since f is a real analytic function, it is enough to show that at least *one* substitution of real numbers satisfies $f(A) \neq 0$ to conclude that its zeroes form a set with measure zero. Hence, an $n \times m$ matrix A satisfies (2) (outside a set of measure zero) provided (6) holds for a value of γ for *some* distribution.

It so happens that a similar statement applies to sets of vectors with a stable sparse representation. As in [9, Sec. IV], consider the "symbolic" dataset $Y = \{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ generated by indeterminate A and k-sparse $\mathbf{a}_1, \ldots, \mathbf{a}_N$.

Theorem 2. *Fix $n, m, k < m$. There is a polynomial in the entries of A and the $\mathbf{a}_i$ with the following property: if real numbers are substituted for the indeterminates such that for every interval of length k in some cyclic order on $[m]$ at least $(k-1)\binom{m}{k} + 1$ of the resulting vectors $\mathbf{a}_i$ are supported on that interval, and the polynomial evaluates to a nonzero number, then Y has a stable k-sparse representation in $\mathbb{R}^m$. In particular, either no substitutions impart to Y this property or all but a Borel set of measure zero do.*

Corollary 2. *Fix n, m, and k satisfying (6) for a value of γ associated to any particular distribution (e.g., that with the smallest known γ), and let the entries of the matrix $A \in \mathbb{R}^{n\times m}$ and k-sparse vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ be drawn independently from probability measures absolutely continuous with respect to the standard Borel measure μ. If at least $(k-1)\binom{m}{k} + 1$ of the vectors $\mathbf{a}_i$ are supported on each interval of length k in some cyclic order on $[m]$, then Y has a stable k-sparse representation in $\mathbb{R}^m$ with probability one.*

An alternative argument made in [9] shows that if $k+1$ random $\mathbf{a}_i$ are drawn for *each* support in $\binom{[m]}{k}$, then Y has a unique k-sparse representation in $\mathbb{R}^m$ (up to permutation-scaling ambiguity) with probability one. This representation can now be said to be stable as well.

We note furthermore that our result in the deterministic case (Thm. 1) accounts for *worst-case* noise. However, for fixed sparsity k, the larger the ambient dimension n of the data, the smaller the probability that the noise points in a direction confusing signals generated by k columns of A. Therefore, for a given distribution, the "effective" noise might be much smaller, with the original dictionary and sparse codes being identifiable for better constants with high probability.

We next address the case when only an upper bound m' on the latent dimension m is known. To do so, we must make the additional assumption that B satisfies (2).

Theorem 3. *Let Y be defined as in the statement of Thm. 1. There exists a constant $C > 0$ for which the following holds for all $\varepsilon < \frac{L_2(A)}{\sqrt{2}}C^{-1}$ and any $m' > m$. If a matrix $B \in \mathbb{R}^{n \times m'}$ satisfies (2) and k-sparse $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^{m'}$ are such that $|A\mathbf{a}_i - B\mathbf{b}_i|_2 \leq \varepsilon$ for all $i \in [N]$ then (4) and (5) hold for some $n \times m$ submatrix of B and corresponding subvectors of the $\mathbf{b}_i$, respectively.*

In other words, the columns of B contain (up to noise, after appropriate scaling) the columns of the original dictionary A. Similarly, the $\mathbf{b}_i$ contain the original codes $\mathbf{a}_i$. The constant C here is expression (29) from the proof of Thm. 3.

III. Proof of Theorem 1

Before proving Thm. 1, we briefly outline our main tools, which include general notions of angle (Def. 2) and distance (Def. 4) between subspaces as well as a (stable) uniqueness result in matrix analysis (Lem. 1). Let $\binom{[m]}{k}$ be all subsets of $[m]$ of size k, and let $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ be the $\mathbb{R}$-linear span of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$. Given $S \subseteq [m]$ and $M \in \mathbb{R}^{n \times m}$, let M_S be the submatrix with columns M_j for $j \in S$, which will also denote its column span when appropriate.

Definition 2. *The **Friedrichs angle** $\theta_F = \theta_F(U, V) \in [0, \frac{\pi}{2}]$ between subspaces $U, V \subseteq \mathbb{R}^n$ is defined in terms of its cosine:*

$$\cos\theta_F := \max\left\{\langle u, v\rangle : \substack{u \in U \cap (U\cap V)^\perp \cap \mathcal{B} \\ v \in V \cap (U \cap V)^\perp \cap \mathcal{B}}\right\},$$

where $\mathcal{B} = \{x : |x|_2 \leq 1\}$ is the unit ℓ_2-ball in $\mathbb{R}^n$ [13].

For example, when $n = 3$ and $k = 1$, this is simply the angle between vectors; and for $k = 2$, it is the angle between the normal vectors of two planes. In higher dimensions, the Friedrichs angle is one out of a set of *principal* (or *canonical* or *Jordan*) angles between subspaces that are invariant to orthogonal transformations. These angles are all zero if and only if one subspace is a subset of the other; otherwise, the Friedrichs angle is the smallest nonzero such angle.

The next quantity is based on one used in [13] to analyze the convergence of the alternating projections algorithm for projecting a point onto the intersection of a set of subspaces.

Definition 3. *Fix $A \in \mathbb{R}^{n \times m}$ and $k < m$. Setting $\phi_1(A) := 1$, define for $k \geq 2$:*

$$\phi_k(A) := \min_{S_1, \ldots, S_k \in \binom{[m]}{k}} 1 - \xi(A_{S_1}, \ldots, A_{S_k}),$$

where for any set $\mathcal{V} = \{V_1, \ldots, V_k\}$ of subspaces of $\mathbb{R}^m$,

$$\xi(\mathcal{V}) := \min_{\sigma \in \mathfrak{S}_k}\left(1 - \prod_{i=1}^{k-1}\sin^2\theta_F\left(V_{\sigma(i)}, \cap_{j=i+1}^k V_j\right)\right)^{1/2},$$

and $\mathfrak{S}_k$ are the permutations (bijections) on k elements.

We are now in a position to state explicitly the constant C referred to in Thm. 1. Letting T be the set of supports on which the $\mathbf{a}_i$ are supported (intervals of length k in some cyclic ordering of $[m]$), X the $m \times N$ matrix with columns $\mathbf{a}_i$, and $I(S) := \{i : S = \text{supp}(\mathbf{a}_i)\}$, we have:

$$C = \left(\frac{\sqrt{k^3}}{\phi_k(A)}\right)\frac{\max_{j\in[m]}|A_j|_2}{\min_{S\in T} L_k(AX_{I(S)})}. \tag{7}$$

Remark 2. *We can be sure that $\min_{S\in T} L_k(AX_{I(S)}) > 0$ so that C is well-defined since $\phi_k(A) = 0$ only when $\text{Span}(A_{S_1}) \supseteq \text{Span}(A_{S_2}) \cap \cdots \cap \text{Span}(A_{S_k})$ for some $S_1, \ldots, S_k \in \binom{[m]}{k}$, which would be in violation of (2).*

Definition 4. *Let U, V be subspaces of $\mathbb{R}^m$ and let $d(u, V) := \inf\{|u - v|_2 : v \in V\} = |u - \Pi_V u|_2$, where Π_V is the orthogonal projection operator onto subspace V. The **gap metric** Θ is defined as [14]:*

$$\Theta(U, V) := \max\left(\sup_{\substack{u \in U \\ |u|_2 = 1}} d(u, V), \sup_{\substack{v \in V \\ |v|_2 = 1}} d(v, U)\right).$$

In fact, $\Theta(U, V)$ is equal to the sine of the largest Jordan angle between U and V.

We now state our uniqueness result in matrix analysis, generalizing [9, Lem. 1] to the noisy case.

Lemma 1 (Main Lemma)**.** *Fix n, m, $k < m$, and let T be the set of intervals of length k in some cyclic ordering of $[m]$. Let $A, B \in \mathbb{R}^{n \times m}$ and suppose that A satisfies the spark condition (2) and has maximum column ℓ_2-norm ρ. If there exists a map $\pi : T \to \binom{[m]}{k}$ and some $\delta < \frac{L_2(A)}{\sqrt{2}}$ such that:*

$$\Theta(A_S, B_{\pi(S)}) \leq \frac{\phi_k(A)}{\rho k}\delta, \quad \textit{for all } S \in T, \tag{8}$$

then there exist a permutation matrix $P \in \mathbb{R}^{m \times m}$ and an invertible diagonal matrix $D \in \mathbb{R}^{m \times m}$ with

$$|A_j - (BPD)_j|_2 \leq \delta, \quad \textit{for } j \in [m]. \tag{9}$$

We will also use the following useful facts about the distance d from Def. 4. The first,

$$\dim(W) = \dim(V) \implies \sup_{\substack{v \in V \\ |v|_2 = 1}} d(v, W) = \sup_{\substack{w \in W \\ |w|_2 = 1}} d(w, V), \tag{10}$$

can be found in [15, Lem. 3.3]. The second is:

Lemma 2. *If U, V are subspaces of $\mathbb{R}^m$, then*

$$d(u, V) < |u|_2, \quad u \in U \setminus \{0\} \implies \dim(U) \leq \dim(V).$$

Proof of Lemma 2. We prove the contrapositive. If $\dim(U) > \dim(V)$, then a dimension argument ($\dim U + \dim V^\perp > m$) gives a nonzero $u \in U \cap V^\perp$. In particular, we have $|u - v|_2^2 = |u|_2^2 + |v|_2^2 \geq |u|_2^2$ for $v \in V$, and thus $d(u, V) \geq |u|_2$. ☐

Finally, we will often make use of the following basic fact:

$$|\mathbf{x}|_1 \leq \sqrt{k}|\mathbf{x}|_2, \quad \text{for } k\text{-sparse } \mathbf{x} \in \mathbb{R}^m. \tag{11}$$

Let us first prove Thm. 1 for the simple case when $k = 1$. Fix $A \in \mathbb{R}^{n\times m}$ satisfying (2), and let $\mathbf{a}_j = c_j\mathbf{e}_j$ for $c_j \in \mathbb{R}\setminus\{0\}$, $j \in [m]$. By (7), we have:

$$C = \sqrt{k^3}\left(\frac{\max_{i\in[m]}|A_i|_2}{\min_{j\in[m]}|c_jA_j|_2}\right) \geq \left(\min_{j\in[m]}|c_j|\right)^{-1}. \quad (12)$$

Suppose that for some $B \in \mathbb{R}^{n\times m}$ and 1-sparse $\mathbf{b}_i \in \mathbb{R}^m$ we have $|A\mathbf{a}_i - B\mathbf{b}_i|_2 \leq \varepsilon < \frac{L_2(A)}{\sqrt{2}}C^{-1}$ for $i \in [m]$. Since the $\mathbf{b}_i$ are 1-sparse, there must exist $c'_1, \ldots, c'_m \in \mathbb{R}$ and some map $\pi : [m] \to [m]$ such that:

$$|c_jA_j - c'_jB_{\pi(j)}|_2 \leq \varepsilon, \quad \text{for} \quad j \in [m]. \quad (13)$$

Note that $c'_j \neq 0$ for all i since otherwise (by definition of $L_2(A)$), we would have $|c_jA_j|_2 < \min_{\ell\in[m]}|c_\ell A_\ell|_2$.

We now show that π is necessarily injective (and thus is a permutation). Suppose that $\pi(i) = \pi(j) = \ell$ for some $i \neq j$ and $\ell \in [m]$. Then, $|c_iA_i - c'_iB_\ell|_2 \leq \varepsilon$ and $|c_jA_j - c'_jB_\ell|_2 \leq \varepsilon$. Scaling and summing these inequalities by $|c'_j|$ and $|c'_i|$, respectively, and applying the triangle inequality, we have:

$$\begin{aligned}(|c'_i| + |c'_j|)\varepsilon &\geq |A(c'_jc_i\mathbf{e}_i - c'_ic_j\mathbf{e}_j)|_2 \\ &\geq \frac{L_2(A)}{\sqrt{2}}\left(|c'_j| + |c'_i|\right)\min_{\ell\in[m]}|c_\ell|, \quad (14)\end{aligned}$$

where the last inequality follows from the definition of $L_2(A)$ and (11). Since (14) contradicts (12) and our upper bound on ε, the map π is injective. Letting $P = \left(\mathbf{e}_{\pi(1)}\cdots\mathbf{e}_{\pi(m)}\right)$ and $D = \text{diag}(\frac{c'_1}{c_1}, \ldots, \frac{c'_m}{c_m})$, we see that (13) becomes for $i \in [m]$:

$$|A_i - (BPD)_i|_2 = |A_i - \frac{c'_i}{c_i}B_{\pi(i)}|_2 \leq \frac{\varepsilon}{|c_i|} \leq C\varepsilon. \quad (15)$$

Remark 3. *It is enough to know* (15) *to bound* $|\mathbf{a}_i - D^{-1}P^{-1}\mathbf{b}_i|_1$ *as well. Specifically, bound* (5) *always follows from* (4) *when* $\varepsilon < \varepsilon_0 := \frac{L_{2k}(A)}{\sqrt{2k}}C^{-1}$. *To see why, note that for all* $2k$*-sparse* $\mathbf{x} \in \mathbb{R}^m$, *we have* $|(A - BPD)\mathbf{x}|_2 \leq C\varepsilon|\mathbf{x}|_1 \leq C\varepsilon\sqrt{2k}|\mathbf{x}|_2$, *by the triangle inequality. Thus,*

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

where in the last inequality we drop the absolute value since $\varepsilon < \varepsilon_0$. *Hence,* $L_{2k}(BPD) \geq L_{2k}(A)(1 - \varepsilon/\varepsilon_0) > 0$ *and:*

$$\begin{aligned}|D^{-1}P^\top\mathbf{b}_i - \mathbf{a}_i|_1 &\leq \sqrt{2k}|\mathbf{a}_i - D^{-1}P^\top\mathbf{b}_i|_2 \\ &\leq \frac{\sqrt{2k}}{L_{2k}(BPD)}|BPD(\mathbf{a}_i - D^{-1}P^\top\mathbf{b}_i)|_2 \\ &\leq \frac{\varepsilon\sqrt{2k}}{L_{2k}(BPD)}(1 + C|\mathbf{a}_i|_1) \\ &\leq \frac{\varepsilon}{\varepsilon_0 - \varepsilon}\left(C^{-1} + |\mathbf{a}_i|_1\right).\end{aligned}$$

It remains to show that (4) with C given in (7) follows from $\varepsilon < \frac{L_2(A)}{\sqrt{2}}C^{-1}$ for $k > 1$. Our main tool is Lem. 1.

Proof of Thm. 1. Let T be the set of intervals of length k in the given cyclic order of $[m]$. From above, we may assume that $k > 1$. Fix $N = m(k-1)\binom{m}{k} + m$ vectors in $\mathbb{R}^k$ as in the statement of the theorem. Fix $A \in \mathbb{R}^{n\times m}$ satisfying (2). We claim that $Y = \{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ has a stable k-sparse representation in $\mathbb{R}^m$. Suppose that for some $B \in \mathbb{R}^{n\times m}$ there exist k-sparse $\mathbf{b}_i \in \mathbb{R}^m$ such that $|A\mathbf{a}_i - B\mathbf{b}_i|_2 \leq \varepsilon$ for all $i \in [N]$. Since there are $(k-1)\binom{m}{k} + 1$ vectors $\mathbf{a}_i$ with a given support $S \in T$, the pigeon-hole principle implies that there exists some $S' \in \binom{[m]}{k}$ and some set of indices $J(S)$ of cardinality k such that all $\mathbf{a}_i$ and $\mathbf{b}_i$ with $i \in J(S)$ have supports S and S', respectively.

Let X and X' be the $m \times N$ matrices with columns $\mathbf{a}_i$ and $\mathbf{b}_i$, respectively. It follows from the general linear position of the $\mathbf{a}_i$ and the linear independence of every k columns of A that the columns of the $n \times k$ matrix $AX_{J(S)}$ are linearly independent, i.e. $L(AX_{J(S)}) > 0$, and therefore form a basis for $\text{Span}\{A_S\}$. Fixing $\mathbf{y} \in \text{Span}\{A_S\}$, there then exists a unique $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^k$ such that $\mathbf{y} = AX_{J(S)}\mathbf{c}$. Letting $\mathbf{y}' = BX'_{J(S)}\mathbf{c}$, which is in $\text{Span}\{B_{S'}\}$, we have:

$$\begin{aligned}|\mathbf{y} - \mathbf{y}'|_2 &= |\sum_{i=1}^k c_i(AX_{J(S)} - BX'_{J(S)})_i|_2 \leq \varepsilon\sum_{i=1}^k|c_i| \\ &\leq \varepsilon\sqrt{k}|\mathbf{c}|_2 \leq \frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})}|AX_{J(S)}\mathbf{c}|_2 = \frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})}|\mathbf{y}|_2.\end{aligned}$$

Hence,

$$\sup_{\substack{\mathbf{y}\in\text{Span}\{A_S\}\\|\mathbf{y}|_2=1}} d(\mathbf{y}, B_{S'}) \leq \frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})}. \quad (16)$$

We now show that (4) follows if $\varepsilon < \frac{L_2(A)}{\sqrt{2}}C^{-1}$, with C as defined in (7). In this case, we can bound the RHS of (16) as follows. Letting $\rho = \max_{j\in[m]}|A_j|_2$ and $I(S) = \{i : \text{supp}(\mathbf{a}_i) = S\}$, we have:

$$\begin{aligned}\frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})} &< \frac{\phi_k(A)L_2(A)}{\rho k\sqrt{2}}\left(\frac{\min_{S\in T}L_k(AX_{I(S)})}{L(AX_{J(S)})}\right) \\ &\leq \frac{\phi_k(A)}{\rho k}\left(\frac{L_2(A)}{\sqrt{2}}\right). \quad (17)\end{aligned}$$

Since $L_2(A) \leq \rho\sqrt{2}$ and $\phi_k(A) \leq 1$, we have that the RHS of (16) is strictly less than one. It follows by Lem. 2 that $\dim(\text{Span}\{B_{S'}\}) \geq \dim(\text{Span}\{A_S\}) = k$ (since every k columns of A are linearly independent). Since $|S'| = k$, we have $\dim(\text{Span}\{B_{S'}\}) \leq k$; hence, $\dim(\text{Span}\{B_{S'}\}) = \dim(\text{Span}\{A_S\})$. Recalling (10), we see the association $S \mapsto S'$ thus defines a map $\pi : T \to \binom{[m]}{k}$ satisfying

$$\Theta(A_S, B_{\pi(S)}) \leq \frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})} \quad \text{for } S \in T. \quad (18)$$

From (17) and (18) we see that the inequality $\Theta(A_S, B_{\pi(S)}) \leq \frac{\phi_k(A)}{\rho k}\delta$ is satisfied for $\delta < \frac{L_2(A)}{\sqrt{2}}$ by setting $\delta = \frac{\rho k}{\phi_k(A)}\left(\frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})}\right)$ (see Rem. 2 for why $\phi_k(A) \neq 0$). We therefore satisfy (8) for

$$\delta = \left(\frac{\varepsilon\sqrt{k^3}}{\phi_k(A)}\right)\frac{\max_{j\in[m]}|A_j|_2}{\min_{S\in T}L_k(AX_{I(S)})} = C\varepsilon.$$

It follows by Lem. 1 that there exist a permutation matrix $P \in \mathbb{R}^{m\times m}$ and invertible diagonal matrix $D \in \mathbb{R}^{m\times m}$ such that $|A_j - (BPD)_j|_2 \leq C\varepsilon$ for all $j \in [m]$. That (5) now follows from this result is contained in Rem. 3. $\square$

IV. Discussion

In this note, we generalized recent results [9] on the uniqueness of solutions to Problem 1 in the noiseless case to the case of noisy measurements, while also reducing the number of required samples from $N = k\binom{m}{k}^2$ to $N = m(k-1)\binom{m}{k} + m$. Surprisingly, almost all $n \times m$ dictionaries satisfying the standard assumption (6) from compressed sensing (CS) are identifiable from N generic noisy k-sparse linear combinations of their elements, up to an error linear in the noise. Moreover, only an upper bound on the number of dictionary elements need be taken as given if solutions are constrained to satisfy (2). We note that hese results extend trivially to the case where certain point-wise injective nonlinearities are applied to the data. We close by outlining four diverse application areas.

Blind Source Separation. Our results provide theoretical grounding for the application of sparse coding to inverse problems ("sparse component analysis"), wherein the linear model (1) is assumed to describe some truth about the data (e.g., the position of a rat on a linear track [16]) and the goal is to infer the generating dictionary and sparse codes from noisy measurements. In this regard, it would be useful to determine for general (m, n, k) the best possible dependence of ε on δ_1, δ_2 (see Def. 1) as well as the minimal requirements on the number and diversity of generating codes. We encourage researchers to extend our results and find tight dependencies on all parameters.

Smoothed Analysis. The main concept in smoothed analysis [17] is that certain algorithms having exponential worst-case behavior are, nonetheless, efficient if certain (typically, measure zero in the continuous case and with "low probability" in the discrete case) pathological input sets are avoided. Our results imply that if there is an efficient "smoothed" algorithm for solving Problem 1 given enough samples, then for generic inputs this algorithm determines the unique original solution. We note that avoiding "bad" (NP-hard) sets of inputs is a necessary technicality for dictionary learning [18], [19].

Neural Communication Theory. In [20] and [21], it was posited that sparse features of natural data passed through a communication bottleneck in the brain using random projections could be decoded, unsupervised, via sparse coding. A necessary condition for this theory to work is that the sparse coding problem has a unique solution. This was already verified in the case of data sampled without noise. Our work extends this theory to the more realistic case of sampling error.

Engineering. Several groups have found ways to utilize CS for signal processing tasks, such as MRI analysis [22], image compression [23] and, more recently, the design of an ultrafast camera [24]. Given such effective uses of classical CS, it is only a matter of time before these systems utilize sparse coding algorithms to encode and process data. In this case, guarantees such as those offered by our theorems allow any such device to be equivalent to any other (having different initial parameters and data samples) as long as enough data originates from a statistically identical system.

We thank Fritz Sommer for turning our attention to the sparse coding problem, Darren Rhea for sharing early explorations, and Ian Morris for posting identity (10) online.

References

[1] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[2] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: algorithms and applications," *Access, IEEE*, vol. 3, pp. 490–530, 2015.

[3] B. Olshausen and M. DeWeese, "Applied mathematics: The statistics of style," *Nature*, vol. 463, no. 7284, pp. 1027–1028, 2010.

[4] J. Hadamard, "Sur les problèmes aux dérivées partielles et leur signification physique," *Princeton university bulletin*, vol. 13, no. 49-52, p. 28, 1902.

[5] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere I: Overview and the geometric picture," *Information Theory, IEEE Transactions on*, 2016.

[6] Y. Li, A. Cichocki, and S.-I. Amari, "Analysis of sparse representation and blind source separation," *Neural computation*, vol. 16, no. 6, pp. 1193–1234, 2004.

[7] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, pp. 992–996, 2005.

[8] M. Aharon, M. Elad, and A. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.

[9] C. Hillar and F. Sommer, "When can dictionary learning uniquely recover sparse data from subsamples?" *Information Theory, IEEE Transactions on*, vol. 61, no. 11, pp. 6290–6297, 2015.

[10] Y. Li, K. Lee, and Y. Bresler, "A unified framework for identifiability analysis in bilinear inverse problems with applications to subspace and sparsity models," *arXiv preprint arXiv:1501.06120*, 2015.

[11] J. Grcar, "A matrix lower bound," *Linear Algebra and its Applications*, vol. 433, no. 1, pp. 203–220, 2010.

[12] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

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

[14] N. Akhiezer and I. Glazman, *Theory of linear operators in Hilbert space*. Courier Corporation, 2013.

[15] I. Morris, "A rapidly-converging lower bound for the joint spectral radius via multiplicative ergodic theory," *Advances in Mathematics*, vol. 225, no. 6, pp. 3425–3445, 2010.

[16] G. Agarwal, I. Stevenson, A. Berényi, K. Mizuseki, G. Buzsáki, and F. Sommer, "Spatially distributed local fields in the hippocampus encode rat position," *Science*, vol. 344, no. 6184, pp. 626–630, 2014.

[17] D. Spielman and S.-H. Teng, "Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time," *Journal of the ACM (JACM)*, vol. 51, no. 3, pp. 385–463, 2004.

[18] M. Razaviyayn, H.-W. Tseng, and Z.-Q. Luo, "Computational intractability of dictionary learning for sparse representation," *arXiv preprint arXiv:1511.01776*, 2015.

[19] A. Tillmann, "On the computational intractability of exact and approximate dictionary learning," *Signal Processing Letters, IEEE*, vol. 22, no. 1, pp. 45–49, 2015.

[20] W. Coulter, C. Hillar, G. Isley, and F. Sommer, "Adaptive compressed sensing – a new class of self-organizing coding models for neuroscience," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5494–5497.

[21] G. Isely, C. Hillar, and F. Sommer, "Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication," in *Advances in neural information processing systems*, 2010, pp. 910–918.

[22] M. Lustig, D. Donoho, J. Santos, and J. Pauly, "Compressed sensing mri," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[23] M. Duarte, M. Davenport, D. Takbar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, March 2008.

[24] L. Gao, J. Liang, C. Li, and L. Wang, "Single-shot compressed ultrafast photography at one hundred billion frames per second," *Nature*, vol. 516, no. 7529, pp. 74–77, 2014.

Charles J. Garfinkle : B.S. Physics, B.S. Chemistry, McGill University, 2010. Currently, Ph.D. candidate, Neuroscience, University of California, Berkeley.

Christopher J. Hillar: B.S. Mathematics, B.S. Computer Science, Yale University, 2000; Ph.D., Mathematics, University of California, Berkeley, CA 2005. Currently, Redwood Center for Theoretical Neuroscience, UCB.