Sparse coding is well-posed for diverse enough samples

Charles J. Garfinkle, Christopher J. Hillar

***Abstract*—Sparse coding or sparse dictionary learning methods have exposed underlying sparse structure in many kinds of natural data. Here, we generalize previous results guaranteeing when the learned dictionary and sparse codes are unique up to inherent permutation and scaling ambiguities [?]. We show that these solutions are robust to the addition of measurement noise provided the data samples are sufficiently diverse. Central to our proofs is a useful lemma in combinatorial matrix theory which allows us to derive bounds on the number of samples necessary to guarantee uniqueness. We also provide probabilistic extensions to our robust identifiability theorem and an extension to the case where only an upper bound on the number of dictionary elements is known a priori. Our results help to inform the interpretation of sparse structure learned from data; whenever the conditions to one of our robust identifiability theorems are met, any sparsity-constrained algorithm that succeeds in approximately reconstructing the data well enough recovers the original dictionary and sparse codes up to an error commensurate with the noise. We discuss applications of this result to smoothed analysis, communication theory, and applied physics and engineering.**

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

I. Introduction

SPARSE coding or dictionary learning algorithms learn to represent each signal in a set $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \subset \mathbb{R}^n$ as a linear combination of few vectors drawn from an inferred *dictionary* $\mathcal{D} \subset \mathbb{R}^n$, where $n \leq |\mathcal{D}| \ll N$ []. One of the domains of application of these algorithms has been in solving inverse problems, wherein the dictionary elements and reconstruction coefficients are interpreted as real physical parameters to be identified from the data []. A proper interpretation of these results demands, however, a characterization of the set of degenerate solutions and their stability with respect to noise.

In this note, we provide conditions guaranteeing uniqueness and stability of solutions to the sparse coding problem. The term 'uniqueness' demands clarification in this context, however, since given any dictionary we can generate a distinct but equally expressive (through linear combination of its elements) dictionary by scaling its elements by some non-zero factors. One can therefore take 'uniqueness' to mean degenerate only up to this ambiguity intrinsic to the statement of the problem. Alternatively, rather than discuss the uniqueness of dictionaries whose elements are vectors combined linearly to reconstruct the data, we can speak of sets of lines through the origin which are combined via the subspace sum $U + V := \{w : w = u + v, u \in U, v \in V\}$. These sets of lines belong to the projective space $P^{n-1}(\mathbb{R})$ defined by identifying all points in $\mathbb{R}^n \setminus \{0\}$ which are non-zero scalings of one another.

The research of Garfinkle and Hillar was conducted while at the Redwood Center for Theoretical Neuroscience, Berkeley, CA, USA; e-mails: cjg@berkeley.edu, chillar@msri.org. Hillar was supported, in part, by National Science Foundation grants IIS-1219212 and IIS-1219199.

Definition 1. *Fix a dataset $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \subset \mathbb{R}^n$. A (m, k)-**scaffolding** of Y is a set $\mathcal{U} = \{U_1, \ldots, U_m\}$ of lines through the origin which satisfies for all $i = 1, \ldots, N$,*

$$\mathbf{y}_i \in U_{j_1} + \ldots + U_{j_k}$$

*for some set of indices $\{j_1, \ldots, j_k\}$ on which we say $\mathbf{y}_i$ is **supported**.*

Question 1. *When does Y have a unique (m, k)-scaffolding?*

Question 2. *Given a (m, k)-scaffolding for Y, when does each datum in Y have a unique support in this scaffolding?*

This question was first addressed in the theoretical works [] which determined how many samples N in the dataset Y are sufficient to guarantee uniqueness. We extend these analyses by asking what happens if the data are noisy and therefore do not necessarily all lie in the subspace sum but rather within some ε-ball around it. To address this question of *stability*, we introduce the following metric on finite subsets of $P^{n-1}(\mathbb{R})$. Let $\mathcal{U} = \{U_i\}_{i \in \mathcal{I}}$ and $\mathcal{V} = \{V_i\}_{i \in \mathcal{I}}$ both be sets of lines indexed by $\mathcal{I}$. We choose:

$$d(\mathcal{U}, \mathcal{V}) = \min_{\sigma \in \mathfrak{S}_{|\mathcal{I}|}} \max_{i \in \mathcal{I}} d'(U_i, V_{\sigma(i)})$$

where $\mathfrak{S}_m$ is the set of permutations on m elements and d' is any metric on $P^{n-1}(\mathbb{R})$ (we will specify one later). [Prove this is a metric. Min-Max Theorem and Compactness?] This measures the largest distance (as given by d') between corresponding lines in the two sets, when the lines have been put into a one-to-one correspondence yielding the smallest possible such value. Let $\mathcal{B}_\varepsilon(\mathbf{X})$ denote the ε-neighbourhood of the set X.

Definition 2. *We say a (m, k)-scaffolding of a dataset is **stable** if for every $\delta \geq 0$ there exists a nonnegative $\varepsilon = \varepsilon(\delta)$ (with $\varepsilon > 0$ when $\delta > 0$) such that all set of lines $\mathcal{V} = \{V_1, \ldots, V_m\}$ satisfying*

$$\mathbf{y}_i \in \mathcal{B}_\varepsilon(V_{j_1} + \ldots + V_{j_k})$$

for some set of indices $\{j_1, \ldots, j_k\}$ satisfy:

$$d(\mathcal{U}, \mathcal{V}) \leq \delta. \tag{1}$$

Question 3. *When does Y have a stable (m, k)-scaffolding?*

[When is the support stable?]

It is easy to see how a solution to Problem **??** directly informs the interpretation of any solution to Problem **??**. Suppose a set of vectors has a robustly identifiable k-sparse representation in $\mathbb{R}^m$ and that Y is a sequence of measurements, each sampled from an ε-ball drawn around these vectors, where $\varepsilon(\delta)$ is defined as in Def. 2 for some fixed $\delta \geq 0$. Then by the triangle inequality, any solution to Problem **??** necessarily satisfies (1). That is, the solution to (**??**) is uniquely determined up to an error commensurate with measurement noise.

The main finding of this work is that so long as no two k-element subsets of dictionary elements span the same k-dimensional subspace, then the k-sparse representation of Y in Problem **??** is unique and stable provided the data are sufficiently diverse. [This is equivalent to the spark condition.] Note that this condition is necessary if we wish for the unique decomposition to be a *direct* sum of subspaces, in particular.

Before stating this result more precisely, we explain how the spark condition relates to the *lower bound* [**?**] of A, which is the largest number α such that $|A\mathbf{x}|_2 > \alpha|\mathbf{x}|_2$ for all $\mathbf{x}$. A straightforward compactness argument shows that every injective linear map has a nonzero lower bound; hence, if A satisfies the spark condition then every submatrix formed from $2k$ of its columns has a nonzero lower bound. We therefore define the following domain-restricted lower bound of a matrix A:

$$L_k(A) := \max\{\alpha : |A\mathbf{s}|_2 \geq \alpha|\mathbf{s}|_2 \text{ for all } k\text{-sparse } \mathbf{s} \in \mathbb{R}^m\}. \tag{2}$$

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

Recall that a *cyclic order* on $[m] := \{1, \ldots, m\}$ is an arrangement of $[m]$ in a circular necklace, and that an *interval* is a contiguous sequence of such elements. The following is our main result.

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

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

$$|A_j - (BPD)_j|_2 \leq C\varepsilon, \quad \text{for all } j \in [m] \tag{3}$$

for some permutation matrix $P \in \mathbb{R}^{m\times m}$ and invertible diagonal matrix $D \in \mathbb{R}^{m\times m}$. If, moreover, we have $\varepsilon < \varepsilon_0$, where $\varepsilon_0 = \frac{L_{2k}(A)}{\sqrt{2k}}C^{-1}$, then B also satisfies the spark condition and

$$|\mathbf{a}_i - D^{-1}P^{-1}\mathbf{b}_i|_1 \leq \frac{\varepsilon}{\varepsilon_0 - \varepsilon}\left(C^{-1} + |\mathbf{a}_i|_1\right), \quad \text{for all } i \in [N]. \tag{4}$$

[*** Should remark that for small enough error the support is recovered. See Donoho paper? So inference is stable for small enough error! There is no multiplicity of alternative models that also happen to work. Regression models have a fixed A (i.e. the values of the variables we are regressing on). Here, in this model, we don't specify the A, instead we learn what the best regressor variables would be if they existed. And we want to combine regressors for different problems so as to save resources. Thought of all this reading section 8 of the Breiman paper on data models vs. algorithmic models.***]

An important consequence of this result is that for sufficiently small reconstruction error, the original dictionary and k-sparse vectors are determined up to a commensurate error and permutation-scaling ambiguity. In terms of Def. 2, for fixed $\delta_1, \delta_2 \geq 0$ we have that (**??**) implies (1) for any $\varepsilon \in [0, \varepsilon_0)$ satisfying:

$$\varepsilon \leq \min\left(\delta_1 C^{-1}, \frac{\delta_2\varepsilon_0}{\delta_2 + C^{-1} + \max_{i\in[N]}|\mathbf{a}_i|_1}\right). \tag{5}$$

The constant C is defined in terms of A and the $\mathbf{a}_i$ in (10), below.

Corollary 1. *Given positive integers n, m, and $k < m$, there exist $N = m(k-1)\binom{m}{k} + m$ vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ with the following property: every matrix $A \in \mathbb{R}^{n\times m}$ satisfying spark condition* (**??**) *generates $Y = \{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ with a robustly identifiable k-sparse representation.*

Our proof of Theorem 1 is a delicate refinement of the arguments in [**?**] to handle measurement error. Here, we also reduce the theoretically required number of samples given by [**?**] from $N = k\binom{m}{k}^2$ to $N = m(k-1)\binom{m}{k} + m$, and we provide guarantees for the case when only an upper bound on the number of columns in A is known (see Theorem 2 below).

Given Theorem 1, it is straightforward to provide probabilistic extensions (Theorems 3 and 4) by drawing on a key result from the recently emergent field of compressed sensing (CS) [**?**], [**?**], [**?**]. [*** Can we just cut right to the chase here with (6)? ***] Briefly, the goal of CS is to recover a signal $\mathbf{x} \in \mathbb{R}^n$ that is sparse enough in some known basis (i.e. $\mathbf{s} = \Psi\mathbf{x}$ is k-sparse for some invertible Ψ) via a stable and efficient reconstruction process after it has been linearly subsampled as $\mathbf{y} = \Phi\mathbf{x}$ by a known compression matrix $\Phi \in \mathbb{R}^{n\times m}$. If the generation matrix $A = \Phi\Psi$ satisfies the spark condition (**??**) then $\mathbf{s}$ is identifiable given $\mathbf{y}$ and the signal $\mathbf{x}$ can then be reconstructed as $\Psi^{-1}\mathbf{s}$. The key result we refer to is that a random[1] compression matrix Φ yields an A satisfying the spark condition with probability one (or "high probability" for discrete variables) provided the dimension n of $\mathbf{y}$ satisfies:

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

where γ is a positive constant dependent on the distribution from which the entries of Φ are drawn. But what if we don't know Φ? Our theorems demonstrate that the matrix A and sparse codes $\mathbf{s}_i$ can still be estimated up to noise and an inherent permutation-scaling ambiguity by examining a sufficiently diverse set of samples $\mathbf{y}_1, \ldots, \mathbf{y}_N$.

We state here the main implications of our probabilistic results. To keep our exposition simple, our statements are

[1] Many ensembles of random matrices work, e.g. [**?**, Sec. §4].

based upon the following elementary construction of *random sparse vectors* (although many ensembles will suffice, e.g. [**?**, Sec. §4]).

Definition 3 (Random k-Sparse Vectors)**.** *Given the support set for its k nonzero entries, a **random draw** of $\mathbf{a}$ is the k-sparse vector with support entries chosen uniformly from the interval $[0,1] \subset \mathbb{R}$, independently.*

Corollary 2. *Suppose m, n, and k satisfy inequality* (6)*. With probability one, a random $n \times m$ generation matrix A satisfies* (**??**)*. Fixing such an A, we have with probability one that a dataset $\{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ generated from a random draw of $N = m(k-1)\binom{m}{k} + m$ k-sparse vectors $\mathbf{a}_i$, consisting of $(k-1)\binom{m}{k} + 1$ samples supported on each interval of length k in some cyclic ordering of $[m]$, has a robustly identifiable k-sparse representation in $\mathbb{R}^m$.*

Corollary 3. *Suppose m, n, and k obey inequality* (6)*. If $N = m(k-1)\binom{m}{k} + m$ randomly drawn vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ are such that for every interval of length k in some cyclic ordering of $[m]$ there are $(k-1)\binom{m}{k} + 1$ vectors $\mathbf{a}_i$ supported on that interval then, with probability one, almost every matrix $A \in \mathbb{R}^{n \times m}$ gives a robustly identifiable $Y = \{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$.*

The organization of the rest of this paper is as follows. In Section II we list additional definitions and key lemmas, including our main tool from combinatorial matrix theory (Lemma 1). We derive Theorem 1 in Section III. In Section IV we state extension of this result to the case where we have only an upper bound on the dimensionality of the sparse codes (or, equivalently, the number of columns in the generating dictionary) and briefly describe their proofs, which are themselves contained in Appendix **??**. In Section V we state and prove Theorems 3 and 4, our probabilistic versions of Theorem 1. The final section is a discussion, and Appendix A contains the proof of Lemma 1.

II. Preliminaries

In this section, we briefly review standard definitions and outline our main tools, which include general notions of angle (Def. 4) and distance (Def. 6) between vector subspaces as well as a robust uniqueness result in combinatorial matrix theory (Lemma 1). Let $\binom{[m]}{k}$ be the collection of subsets of $[m] := \{1, \ldots, m\}$ of cardinality k, and let $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ for real vectors $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$ be the vector space consisting of their $\mathbb{R}$-linear span. For a matrix M, the spectral norm is denoted $\|M\|_2$. Also, given $S \subseteq [m]$ and $M \in \mathbb{R}^{n \times m}$ with columns $\{M_1, \ldots, M_m\}$, we define M_S to be the submatrix with columns $\{M_j : j \in S\}$ and also set $\text{Span}\{M_S\} := \text{Span}\{M_j : j \in S\}$.

Between any pair of subspaces in Euclidean space one can define the following generalized "angle":

Definition 4. *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:*

$$\cos\theta_F := \max\left\{\langle u, v\rangle : u \in U \cap (U \cap V)^\perp \cap \mathcal{B},\ v \in V \cap (U \cap V)^\perp \cap \mathcal{B}\right\}, \tag{7}$$

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

For example, for $n = 3$ 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 which 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 definition we need is based on a quantity derived in [**?**] to describe the convergence of the alternating projections algorithm for projecting a point onto the intersection of a set of subspaces. We use it to bound the distance between a point and the intersection of a set of subspaces given an upper bound on the distance from that point to each individual subspace. The minimization over permutations in (9) below is done only to remove the dependence the convergence result has on the order in which subspaces are inputted to the alternating projections algorithm.

Definition 5. *Fix $A \in \mathbb{R}^{n \times m}$ and $k < m$. Setting $\phi_1(A) := 1$, define for $k \geq 2$:*

$$\phi_k(A) := \min_{S_1, \ldots, S_k \in \binom{[m]}{k}} 1 - \xi(\text{Span}\{A_{S_1}\}, \ldots, \text{Span}\{A_{S_k}\}), \tag{8}$$

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

$$\xi(\mathcal{V}) := \min_{\sigma \in \mathfrak{S}_k} \left(1 - \prod_{i=1}^{k-1} \sin^2\theta_F\left(V_{\sigma(i)}, V_{\sigma(i+1)} \cap \cdots \cap V_{\sigma(k)}\right)\right)^{1/2}. \tag{9}$$

and $\mathfrak{S}_k$ is the set of permutations (i.e. bijections) on k elements.

We are now in a position to state explicitly the constant C referred to in Theorem 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 be the $m \times N$ matrix with columns $\mathbf{a}_i$, let, and $I(S) := \{i : \text{supp}(\mathbf{a}_i) = S\}$, we have:

$$C = \left(\frac{\sqrt{k^3}}{\phi_k(A)}\right) \frac{\max_{j \in [m]} |A_j|_2}{\min_{S \in T} L_k(AX_{I(S)})}, \tag{10}$$

Remark 1. *We can be sure that C is well-defined provided $\min_{S \in T} L_k(AX_{I(S)}) > 0$, since $\phi_k(A) = 0$ only when Span$(A_{S_1}) \supseteq$ Span$(A_{S_2}) \cap \ldots, \cap$Span$(A_{S_k})$ for some $S_1, \ldots, S_k \in \binom{[m]}{k}$, which would be in violation of the spark condition on A.*

Definition 6. *Let U, V be subspaces of $\mathbb{R}^m$ and let $d(u, V) := \inf\{|u - v|_2 : v \in W\} = |u - \Pi_V u|_2$ where Π_V is the orthogonal projection operator onto subspace V. The **gap metric** Θ on subspaces of $\mathbb{R}^m$ is [**?**]:*

$$\Theta(U, V) := \max\left(\sup_{u \in U,\, |u|_2 = 1} d(u, V),\ \sup_{v \in V,\, |v|_2 = 1} d(v, U)\right). \tag{11}$$

Remark 2. *We note that $\Theta(U, V)$ is in fact equal to the sine of the largest Jordan angle between U and V.*

We now state our main result from combinatorial matrix theory, generalizing [?, Lemma 1] to the noisy case.

Lemma 1 (Main Lemma)**.** *Fix positive integers 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 (**??**) with 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*

$$\Theta(\text{Span}\{A_S\}, \text{Span}\{B_{\pi(S)}\}) \le \frac{\phi_k(A)}{\rho k}\delta, \quad \textit{for all } S \in T, \tag{12}$$

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

$$|A_j - (BPD)_j|_2 \le \delta, \quad \textit{for all } j \in [m]. \tag{13}$$

We defer the (technical) proof of this lemma to Appendix A. In words, the result says that the vectors forming the columns of A are nearly identical to those forming the columns of B (up to symmetry) provided that for a special set T of m subspaces, each spanned by k columns of A, there exist k columns of B which span a nearby subspace with respect to the gap metric.

In our proof of robust identifiability, we will also use the following useful facts about the distance d from Def. 6. The first,

$$\dim(W) = \dim(V) \implies \sup_{v\in V, |v|_2=1} d(v, W) = \sup_{w\in W, |w|_2=1} d(w, V), \tag{14}$$

can be found in [?, Lemma 3.3]. The second is:

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

$$d(u, V) < |u|_2 \ \text{ for all } u \in U \setminus \{0\} \implies \dim(U) \le \dim(V). \tag{15}$$

Proof. 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 $|u - v|_2^2 = |u|_2^2 + |v|_2^2 \ge |u|_2^2$ for all $v \in V$, and thus $d(u, V) \ge |u|_2$. □

III. Deterministic Identifiability

Let $\mathbf{e}_1, \ldots, \mathbf{e}_m$ be the standard basis vectors in $\mathbb{R}^m$. Before proving Theorem 1 in full generality, consider when $k = 1$. Fix some $A \in \mathbb{R}^{n\times m}$ satisfying spark condition (**??**) and suppose we have $N = m$ 1-sparse vectors $\mathbf{a}_j = c_j\mathbf{e}_j$ for $c_j \in \mathbb{R}\setminus\{0\}$, $j \in [m]$. By (10) we have:

$$C = \left(\frac{\sqrt{k^3}}{\phi_1(A)}\right)\frac{\max_{j\in[m]}|A_j|_2}{\min_{i\in[m]} L_1(c_iA_i)} = \sqrt{k^3}\left(\frac{\max_{j\in[m]}|A_j|_2}{\min_{i\in[m]}|c_iA_i|_2}\right) \ge \max_{i\in[m]}\frac{1}{|c_i|}, \tag{16}$$

Suppose that for some $B \in \mathbb{R}^{n\times m}$ and 1-sparse $\mathbf{b}_i \in \mathbb{R}^m$ we have $|A\mathbf{a}_i - B\mathbf{b}_i|_2 \le \varepsilon < \frac{L_2(A)}{\sqrt{2}}C^{-1}$ for all $i \in [m]$. 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

$$|c_iA_i - c'_iB_{\pi(i)}|_2 \le \varepsilon \ \text{ for all } i \in [m]. \tag{17}$$

Note that $c'_i \neq 0$ for all i since then otherwise (by definition of $L_2(A)$) we reach the contradiction $|c_iA_i|_2 < \min_{i\in[m]}|c_iA_i|_2$. We will now show that π is necessarily injective (and thus defines a permutation). Suppose that $\pi(i) = \pi(j) = \ell$ for some $i \neq j$ and $\ell \in [m]$. Then, $|c_iA_i - c'_iB_\ell|_2 \le \varepsilon$ and $|c_jA_j - c'_jB_\ell|_2 \le \varepsilon$. Scaling and summing these inequalities by $|c'_j|$ and $|c'_i|$, respectively and then applying the triangle inequality, we have:

$$(|c'_i| + |c'_j|)\varepsilon \ge |A(c'_jc_i\mathbf{e}_i - c'_ic_j\mathbf{e}_j)|_2 \ge \frac{L_2(A)}{\sqrt{2}}\left(|c'_j| + |c'_i|\right)\min_{\ell\in[m]}|c_\ell|, \tag{18}$$

where the last inequality follows from the definition of $L_2(A)$ and the fact that $|\mathbf{x}|_1 \le \sqrt{k}|\mathbf{x}|_2$ for k-sparse $\mathbf{x}$. Since (18) is in contradiction with (16) and our upper bound on ε, it must be that π is in fact 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 (17) becomes

$$|A_i - (BPD)_i|_2 = |A_i - \frac{c'_i}{c_i}B_{\pi(i)}|_2 \le \frac{\varepsilon}{|c_i|} \le C\varepsilon \ \text{ for all } i \in [m]. \tag{19}$$

Remark 3. *Only minor modifications of the above arguments are necessary to prove a generalization of Theorem 1 for the case $k = 1$ where B may have more than m columns (i.e. $B \in \mathbb{R}^{n\times m'}$ and $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^{m'}$ for some $m' \ge m$). In this case, from the injective function $\pi : [m] \to [m']$ we may define a 'partial' permutation matrix $P = \left(\mathbf{e}_{\pi(1)}\cdots\mathbf{e}_{\pi(m)}, \mathbf{0}, \ldots, \mathbf{0}\right) \in \mathbb{R}^{m'\times m'}$ having at most one nonzero entry in each row and column and every one of these nonzero entries being one, and the diagonal matrix $D \in \mathbb{R}^{m'\times m}$ still satisfying $D_{ij} = 0$ whenever $i \neq j$.*

Remark 4. *From* (19) *we can, in general, bound $|\mathbf{a}_i - D^{-1}P^{-1}\mathbf{b}_i|_1$ as well. Specifically, we will show that* (4) *always follows from* (3) *when $\varepsilon < \varepsilon_0$, with $\varepsilon_0 = \frac{L_{2k}}{\sqrt{2k}}C^{-1}$. Note that for all $2k$-sparse $\mathbf{x} \in \mathbb{R}^m$ we have by the triangle inequality:*

$$|(A - BPD)\mathbf{x}|_2 \le C\varepsilon|\mathbf{x}|_1 \le C\varepsilon\sqrt{2k}|\mathbf{x}|_2.$$

Thus,

$$|BPD\mathbf{x}|_2 \ge ||A\mathbf{x}|_2 - |(A - BPD)\mathbf{x}|_2| \ge (L_{2k}(A) - \sqrt{2k}C\varepsilon)|\mathbf{x}|_2,$$

where for the last inequality it was admissible to drop the absolute value since $\varepsilon < \varepsilon_0$. Hence, $L_{2k}(BPD) \ge L_{2k}(A) - C\varepsilon\sqrt{2k} > 0$ and it follows that:

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

Remark 5. *We demonstrate with the following counter-example that for C as defined in* (10) *the condition $\varepsilon < \frac{L_2(A)}{\sqrt{2}} C^{-1}$ is necessary to guarantee in general that* (3) *follows from the remaining assumptions of Theorem 1. Consider the dataset $\mathbf{a}_i = \mathbf{e}_i$ for $i = 1, \ldots, m$ and let A be the identity matrix in $\mathbb{R}^{m \times m}$. Then $L_2(A) = 1$ (we have $|A\mathbf{x}|_2 = |\mathbf{x}|_2$ for all $\mathbf{x} \in \mathbb{R}^m$) and $C = 1$; hence $\frac{L_2(A)}{\sqrt{2}} C^{-1} = 1/\sqrt{2}$. Consider the alternate dictionary $B = \left(\mathbf{0}, \frac{1}{2}(\mathbf{e}_1 + \mathbf{e}_2), \mathbf{e}_3, \ldots, \mathbf{e}_m\right)$ and sparse codes $\mathbf{b}_i = \mathbf{e}_2$ for $i = 1, 2$ and $\mathbf{b}_i = \mathbf{e}_i$ for $i = 3, \ldots, m$. Then $|A\mathbf{a}_i - B\mathbf{b}_i| = 1/\sqrt{2}$ for $i = 1, 2$ (and 0 otherwise). If there were permutation and invertible diagonal matrices $P \in \mathbb{R}^{m \times m}$ and $D \in \mathbb{R}^{m \times m}$ such that $|(A - BPD)\mathbf{e}_i| \le C\varepsilon$ for all $i \in [m]$, then we would reach the contradiction $1 = |P^{-1}\mathbf{e}_1|_2 = |(A - BPD)P^{-1}\mathbf{e}_1|_2 \le 1/\sqrt{2}$.*

It remains to show that (3) with C given in (10) follows from $\varepsilon < \frac{L_2(A)}{\sqrt{2}} C^{-1}$ for $k > 1$. Our main tool for the proof is Lemma 1.

Proof of Theorem 1 and Corollary 1. Let T be the set of intervals of length k in the given cyclic order of $[m]$. From above, we may assume that $k > 1$. The first step is to produce a set of $N = m(k-1)\binom{m}{k} + m$ vectors in $\mathbb{R}^k$ in general linear position (i.e., any k of them are linearly independent). Specifically, let $\gamma_1, \ldots, \gamma_N$ be any distinct numbers. Then the columns of the $k \times N$ matrix $V = (\gamma_j^i)_{i,j=1}^{k,N}$ are in general linear position (since the γ_j are distinct, any $k \times k$ "Vandermonde" sub-determinant is nonzero). Next, form the k-sparse vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ with supports $S \in T$ (partitioning the a_i evenly among these supports so that each support contains $(k-1)\binom{m}{k} + 1$ vectors a_i) by setting the nonzero values of vector $\mathbf{a}_i$ to be those contained in the ith column of V.

Fix $A \in \mathbb{R}^{n \times m}$ satisfying (**??**). We claim that $\{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ has a robustly identifiable 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 $|A\mathbf{a}_i - B\mathbf{b}_i|_2 \le \varepsilon$ 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 Span$\{A_S\}$. Fixing $\mathbf{z} \in$ Span$\{A_S\}$, there then exists a unique $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^k$ such that $\mathbf{z} = AX_{J(S)}\mathbf{c}$. Letting $\mathbf{z}' = BX'_{J(S)}\mathbf{c}$, which is in Span$\{B_{S'}\}$, we have:

$$|\mathbf{z} - \mathbf{z}'|_2 = |\sum_{j=1}^{N} c_i (AX_{J(S)} - BX'_{J(S)})\mathbf{e}_j|_2 \le \varepsilon \sum_{j=1}^{N} |c_j| \le \varepsilon\sqrt{k}|\mathbf{c}|_2 \le \frac{\varepsilon\sqrt{k}}{L(AX_{JS})}|AX_{J(S)}\mathbf{c}|_2 = \frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})}|\mathbf{z}|_2.$$

Hence,

$$\sup_{\substack{\mathbf{z} \in \text{Span}\{A_S\} \\ |\mathbf{z}|_2 = 1}} d(\mathbf{z}, \text{Span}\{B_{S'}\}) \le \frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})}. \tag{20}$$

We now show that (3) follows if $\varepsilon < \frac{L_2(A)}{\sqrt{2}} C^{-1}$, with C as defined in (10). In this case we can bound the RHS of (20) as follows. Letting $\rho = \max_{j \in [m]} |A_j|_2$ and $I(S) = \{i : \text{supp}(\mathbf{a}_i) = S\}$, we have:

$$\frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})} < \frac{\phi_k(A) L_2(A)}{\rho k \sqrt{2}} \left(\frac{\min_{S \in T} L_k(AX_{I(S)})}{L(AX_{J(S)})} \right) \le \frac{\phi_k(A)}{\rho k} \left(\frac{L_2(A)}{\sqrt{2}} \right. \tag{21}$$

Since $L_2(A) \le \rho\sqrt{2}$ and $\phi_k(A) \le 1$, we have that the RHS of (20) is strictly less than one. It follows by Lemma 2 that $\dim(\text{Span}\{B_{S'}\}) \ge \dim(\text{Span}\{A_S\}) = k$ (since every k columns of A are linearly independent). Since $|S'| = k$, we have $\dim(\text{Span}\{B_{S'}\}) \le k$; hence, $\dim(\text{Span}\{B_{S'}\}) = \dim(\text{Span}\{A_S\})$. Recalling (14), we see the association $S \mapsto S'$ thus defines a map $\pi : T \to \binom{[m]}{k}$ satisfying

$$\Theta(\text{Span}\{A_S\}, \text{Span}\{B_{\pi(S)}\}) \le \frac{\varepsilon\sqrt{k}}{L(AX_{J(S)})} \quad \text{for all } S \in T. \tag{22}$$

From (21) and (22) we see that the inequality $\Theta(\text{Span}\{A_S\}, \text{Span}\{B_{\pi(S)}\}) \le \frac{\phi_k(A)}{\rho k}\delta$ 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 Remark 1 for why we can be sure $\phi_k(A) \neq 0$). We therefore satisfy (12) for

$$\delta = \left(\frac{\varepsilon\sqrt{k^3}}{\phi_k(A)} \right) \frac{\max_{j \in [m]} |A_j|_2}{\min_{S \in T} L_k(AX_{I(S)})} = C\varepsilon \tag{23}$$

It follows by Lemma 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 $|A_j - (BPD)_j|_2 \le C\varepsilon$ for all $j \in [m]$. The proof of how (4) follows from this result is contained in Remark 4. □

IV. Unknown representation dimension

In this section we state a version of Theorem 1 and Lemma 1 assuming that B also satisfies the spark condition (in addition to A satisfying the spark condition). With this additional assumption, we can address the issue of recovering $A \in \mathbb{R}^{n \times m}$ and the $\mathbf{a}_i \in \mathbb{R}^m$ when only an upper bound m' on the number m of columns in A is known.

Theorem 2. *Fix positive integers n, m, m', and k with $k < m \le m'$ and fix a cyclic order on $[m]$. If k-sparse $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ are such that for every interval of length k in the cyclic order there are $(k-1)\binom{m'}{k} + 1$ vectors $\mathbf{a}_i$ in general linear position supported on that interval and $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times m'}$ both satisfy spark condition* (**??**) *then there exists a constant $\tilde{C} > 0$ for which the following holds for all $\varepsilon < \frac{L_2(A)}{\sqrt{2}} \tilde{C}^{-1}$. If there exist k-sparse $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^{m'}$ such that $|A\mathbf{a}_i - B\mathbf{b}_i|_2 \le \varepsilon$ for all $i \in [N]$ then:*

$$|A_j - (BPD)_j|_2 \le \tilde{C}\varepsilon \quad \text{for all } j \in [m] \tag{24}$$

for some partial permutation matrix $P \in \mathbb{R}^{m' \times m'}$ (there is at most one nonzero entry in each row and column and these nonzero entries are all one) and diagonal matrix $D \in \mathbb{R}^{m' \times m}$ ($D_{ij} = 0$ whenever $i \neq j$).

In other words, the columns of the learned dictionary B contain (up to noise, and after appropriate scaling) the columns of the original dictionary A. One can then show by manipulations similar to those contained in Remark 4 that the coefficients in each $\mathbf{a}_i$ form (up to noise) a scaled subset of the coefficients in $\mathbf{b}_i$. The constant $\tilde{C}$ is given by:

$$\tilde{C} = \left(\frac{\sqrt{k^3}}{\min(\phi_k(A), \phi_k(B))} \right) \frac{\max_{j \in [m]} |A_j|_2}{\min_{S \in T} L_k(AX_{I(S)})}, \quad (25)$$

where X is the $m \times N$ matrix with columns given by $\mathbf{a}_i$ and $I(S) = \{i : \text{supp}(\mathbf{a}_i) = S\}$.

The proof of Theorem 2 is very similar to the proof of Theorem 1, the difference being that now we establish a map $\pi : [m] \to [m']$ satisfying the requirements of Lemma 3, which we state next, by pigeonholing $(k-1)\binom{m'}{k}+1$ vectors with respect to holes $[m']$. This insures that we can establish a one-to-one correspondence between subspaces spanned by the m' columns of B and nearby subspaces spanned by the m columns of A despite the fact that $m < m'$. By requiring B to also satisfy the spark condition, we remove the dependency of Lemma 1 on Lemma 7 (which requires that $m = m'$), resulting in Lemma 3.

Lemma 3 (Main Lemma for $m < m'$)**.** *Fix positive integers n, m, m', and k where $k < m < m'$, and let T be the set of intervals of length k in some cyclic ordering of $[m]$. Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times m'}$ both satisfy spark condition* (**??**) *with A having 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*

$$\Theta(\text{Span}\{A_S\}, \text{Span}\{B_{\pi(S)}\}) \leq \frac{\delta}{\rho k} \min(\phi_k(A), \phi_k(B)), \quad \textit{for all } S \in T, \quad (26)$$

then there exist a partial permutation matrix $P \in \mathbb{R}^{m' \times m'}$ and a diagonal matrix $D \in \mathbb{R}^{m' \times m}$ such that

$$|A_j - (BPD)_j|_2 \leq \delta, \quad \textit{for all } j \in [m]. \quad (27)$$

We defer the proof of this lemma to Appendix **??**.

V. Probabilistic Identifiability

We next give precise statements of our probabilistic versions of Theorem 1, which apply to random sparse vectors as defined in Definition 3. Our brief proofs rely largely on the following lemma, the proof of which can be found in [**?**, Lemma 3]:

Lemma 4. *Fix positive integers n, m, $k < m$, and a matrix $M \in \mathbb{R}^{n \times m}$ satisfying* (**??**)*. With probability one, $M\mathbf{a}_1, \ldots, M\mathbf{a}_k$ are linearly independent whenever the $\mathbf{a}_i$ are random k-sparse vectors.*

Theorem 3. *Fix positive integers n, m, $k < m$, and $A \in \mathbb{R}^{n \times m}$ satisfying* (**??**)*. If a set of $N = m(k-1)\binom{m}{k} + m$ randomly drawn vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ are such that for each interval of length k in some cyclic order on $[m]$ there are $(k-1)\binom{m}{k} + 1$ vectors $\mathbf{a}_i$ supported on that interval then $\{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ has a robustly identifiable k-sparse representation with probability one.*

Proof. By Lemma 4 (setting M to be the identity matrix), with probability one the $\mathbf{a}_i$ are in general linear position. Apply Theorem 1. □

Note that an *algebraic set* is a solution to a finite set of polynomial equations.

Theorem 4. *Fix positive integers $k < m$ and n . If $N = m(k-1)\binom{m}{k} + m$ randomly drawn vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ are such that for every interval of length k in some cyclic ordering of $[m]$ there are $(k-1)\binom{m}{k} + 1$ vectors $\mathbf{a}_i$ supported on that interval, then with probability one the following holds. There is an algebraic set $Z \subset \mathbb{R}^{n \times m}$ of Lebesgue measure zero with the following property: if $A \notin Z$ then $\{A\mathbf{a}_1, \ldots, A\mathbf{a}_N\}$ has a robustly identifiable k-sparse representation in $\mathbb{R}^m$.*

Proof. By Lemma 4 (setting M to be the identity matrix), with probability one the $\mathbf{a}_i$ are in general linear position. By the same arguments made in the proof of Theorem 3 in [**?**], the set of matrices A that fail to satisfy (**??**) form an algebraic set of measure zero. Apply Theorem 1. □

We note that our results in the deterministic case (Theorem 1) were derived for the *worst case* noise, which (depending on the noise distribution) may have very small probability of occurrence. In particular, for fixed k, the larger the ambient dimension of $\mathbf{y}$, the smaller the probability that randomly generated noise points in a direction which conflates signals generated by distinct k-dimensional subspaces spanned by the columns of A. For a given distribution the "effective" noise may therefore potentially be much smaller, with the original dictionary and sparse codes being identifiable up to a commensurate error with high probability.

VI. Discussion

In this note, we generalize the known uniqueness of solutions to (**??**) in the exact case ($\varepsilon = 0$) to the more realistic case of deterministic noise ($\varepsilon > 0$). Somewhat surprisingly, as long as the standard assumptions from compressed sensing are met, almost every dictionary and sufficient quantity N of sparse codes are uniquely determined up to the error in sampling and the problem's inherent symmetries (uniform relabeling and scaling). To convince the reader of the general usefulness of this result, we elaborate briefly on four diverse application areas.

Sparse Component Analysis (SCA). In the SCA framework, it is assumed that the linear model in (**??**) describes some real physical process and the goal is to infer the true underlying dictionary A and sparse codes $\mathbf{a}_i$ from measurements $\mathbf{y}_i$. Often in such analyses it is implicitly assumed that sparse coding or dictionary learning algorithms have exposed this underlying sparse structure in the data (e.g. the feature-tuned field potentials in [**?**]) as opposed to some artifactual degenerate solution. Our results suggest that given enough data samples the uniqueness of this decomposition is indeed the norm rather than the exception. Regarding this, it would be useful to determine for general k the best possible dependence of ε on δ_1, δ_2 (see Def. 2) as well as the minimum possible sample size N. We encourage the sparse coding community to extend our results and find a tight dependency of all the parameters, both for the sake of theory and practical applications.

Smoothed Analysis. The main idea in smoothed analysis [**?**] is that certain algorithms having bad worst case behavior, nonetheless, are efficient if certain (typically Lebesgue measure zero) pathological input sets are avoided. Our results imply that if there is an efficient "smoothed" algorithm for solving (**??**) then only for a measure zero set of inputs will this algorithm fail to determine the unique original solution; this is a focus of future work. We note that avoiding "bad" (NP-hard) sets of inputs is necessary for dictionary learning [**?**], [**?**].

Communication Theory. In [**?**] and [**?**] it was posited that sparse features of natural data could be passed through a communication bottleneck in the brain using random projections and decoded by unsupervised learning of sparse codes. A necessary condition for this theory to work is that there is a unique solution to the SCA problem. This was verified in the case of data sampled without noise in [**?**]. The present work extends this theory to the more realistic case of sampling noise.

Applied Physics and Engineering. As a final application domain, we consider engineering applications. Several groups have found ways to utilize compressed sensing for signal processing tasks, such as digital image compression [**?**] (the "single-pixel camera") and, more recently, the design of an ultrafast camera [**?**] capable of capturing one hundred billion frames per second. Given such effective uses of classical CS, it is only a matter of time before these systems utilize sparse coding algorithms to code and process data. In this case, guarantees such as the ones offered by our main theorems allow any such device to be compared to any other (having different initial parameters and data samples) as long as the data originates from the same system.

Appendix A
Combinatorial Matrix Theory

In this section, we prove Lemma 1, which is the main ingredient in our proof of Theorem 1. For readers willing to assume a priori that the spark condition holds for B as well as for A, a shorter proof of this case (Lemma 3 from Section IV) is provided in Appendix **??**. This additional assumption simplifies the argument and allows us to extend robust identifiability conditions to the case where only an upper bound on the number of columns m in A is known.

We now prove some auxiliary lemmas before deriving Lemma 1. Given a collection of sets $\mathcal{T}$, we let $\cap\mathcal{T}$ denote their intersection.

Lemma 5. *Let $M \in \mathbb{R}^{n\times m}$. If every $2k$ columns of M are linearly independent, then for any $\mathcal{T} \subseteq \bigcup_{\ell\le k}\binom{[m]}{\ell}$,*

$$\text{Span}\{M_{\cap\mathcal{T}}\} = \bigcap_{S\in\mathcal{T}} \text{Span}\{M_S\}. \tag{28}$$

Proof. By induction, it is enough to prove the lemma when $|\mathcal{T}| = 2$. The proof now follows directly from the assumption. ☐

Lemma 6. *Fix $k \ge 2$. Let $\mathcal{V} = \{V_1,\ldots,V_k\}$ be subspaces of $\mathbb{R}^m$ and let $V = \bigcap\mathcal{V}$. For every $\mathbf{x}\in\mathbb{R}^m$, we have*

$$|\mathbf{x} - \Pi_V\mathbf{x}|_2 \le \frac{1}{1-\xi(\mathcal{V})}\sum_{i=1}^{k}|x - \Pi_{V_i}x|_2, \tag{29}$$

provided $\xi(\mathcal{V}) \ne 1$, where the expression for ξ is given in Def. 5.

Proof. Fix $\mathbf{x}\in\mathbb{R}^m$ and $k \ge 2$. The proof consists of two parts. First, we shall show that

$$|\mathbf{x}-\Pi_V\mathbf{x}|_2 \le \sum_{\ell=1}^{k}|\mathbf{x}-\Pi_{V_\ell}\mathbf{x}|_2 + |\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\mathbf{x}-\Pi_V\mathbf{x}|_2. \tag{30}$$

For each $\ell \in \{2,\ldots,k+1\}$ (when $\ell = k+1$, the product $\Pi_{V_k}\cdots\Pi_{V_\ell}$ is set to I), we have by the triangle inequality and the fact that $\|\Pi_{V_\ell}\|_2 \le 1$ (as Π_{V_ℓ} are projections):

$$|\Pi_{V_k}\cdots\Pi_{V_\ell}\mathbf{x}-\Pi_V\mathbf{x}| \le |\Pi_{V_k}\cdots\Pi_{V_{\ell-1}}\mathbf{x}-\Pi_V\mathbf{x}| + |\mathbf{x}-\Pi_{V_{\ell-1}}\mathbf{x}|. \tag{31}$$

Summing these inequalities over ℓ gives (30).

Next, we show how the result (29) follows from (30) from the following result of [**?**, Theorem 9.33]:

$$|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\mathbf{x} - \Pi_V\mathbf{x}|_2 \le z|\mathbf{x}|_2 \quad \text{for all} \quad \mathbf{x}\in\mathbb{R}^m, \tag{32}$$

where $z = \left[1 - \prod_{\ell=1}^{k-1}(1-z_\ell^2)\right]^{1/2}$ and $z_\ell = \cos\theta_F\left(V_\ell, \cap_{s=\ell+1}^{k}V_s\right)$. To see this, note that

$$|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}(\mathbf{x}-\Pi_V\mathbf{x}) - \Pi_V(\mathbf{x}-\Pi_V\mathbf{x})|_2 = |\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\mathbf{x} \tag{33}$$

since $\Pi_{V_\ell}\Pi_V = \Pi_V$ for all $\ell = 1,\ldots,k$ and $\Pi_V^2 = \Pi_V$. Therefore by (32) and (33), it follows that

$$|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\mathbf{x}-\Pi_V\mathbf{x}|_2 = |\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}(\mathbf{x}-\Pi_V\mathbf{x})-\Pi_V(\mathbf{x}-\Pi$$

Combining this last inequality with (30) and rearranging, we arrive at

$$|\mathbf{x} - \Pi_V\mathbf{x}|_2 \le \frac{1}{1-z}\sum_{i=1}^{k}|\mathbf{x} - \Pi_{V_i}\mathbf{x}|_2. \tag{34}$$

Finally, since the ordering of the subspaces is arbitrary, we can replace z in (34) with $\xi(\mathcal{V})$ to obtain (29). ☐

Lemma 7. *Fix positive integers $k < m$, and let $S_1,\ldots,S_m$ be the set of contiguous length k intervals in some cyclic order of $[m]$. Suppose there exists a map $\pi : T \to \binom{[m]}{k}$ such that*

$$|\bigcap_{i\in J}\pi(S_i)| \le |\bigcap_{i\in J}S_i| \quad \text{for all } J \in \binom{[m]}{k}. \tag{35}$$

Then, $|\pi(S_{j_1})\cap\cdots\cap\pi(S_{j_k})| = 1$ for $j_1,\ldots,j_k$ consecutive modulo m.

Proof. Consider the set $Q_m = \{(r,t) : r\in\pi(S_t), t\in[m]\}$, which has mk elements. By the pigeon-hole principle, there is some $q\in[m]$ and $J\in\binom{[m]}{k}$ such that $(q,j)\in Q_m$ for all $j\in J$. In particular, we have $q\in\cap_{j\in J}\pi(S_j)$ so that from (35) there must be some $p\in[m]$ with $p\in\cap_{j\in J}S_j$. Since $|J| = k$, this is only possible if the elements of $J = \{j_1,\ldots,j_k\}$ are consecutive modulo m, in which case $|\cap_{j\in J} S_j| = 1$. Hence $|\cap_{j\in J} \pi(S_j)| = 1$ as well.

We next consider if any other $t \notin J$ is such that $q \in \pi(S_t)$. Suppose there were such a t; then, we would have $q \in \pi(S_t) \cap \pi(S_{j_1}) \cap \cdots \cap \pi(S_{j_k})$ and (35) would imply that the intersection of every k-element subset of $\{S_t\} \cup \{S_j : j \in J\}$ is nonempty. This would only be possible if $\{t\} \cup J = [m]$, in which case the result then trivially holds since then $q \in \pi(S_j)$ for all $j \in [m]$. Suppose now there exists no such t; then letting $Q_{m-1} \subset Q_m$ be the set of elements of Q_m not having q as a first coordinate, we have $|Q_{m-1}| = (m-1)k$.

By iterating the above arguments we arrive at a partitioning of Q_m into sets $R_i = Q_i \setminus Q_{i-1}$ for $i = 1, \ldots, m$, each having a unique element of $[m]$ as a first coordinate common to all k elements while having second coordinates which form a consecutive set modulo m. In fact, every set of k consecutive integers modulo m is the set of second coordinates of some R_i. This must be the case because for every consecutive set J we have $|\cap_{j\in J} S_j| = 1$, whereas if J is the set of second coordinates for two distinct sets R_i we would have $|\cap_{j\in J} \pi(S_j)| \geq 2$, which violates (35). □

Proof of Lemma 1 (Main Lemma). We assume $k \geq 2$ since the case $k = 1$ was proven at the beginning of Section III. Let $S_1, \ldots, S_m$ be the set of contiguous length k intervals in some cyclic ordering of $[m]$. We begin by proving that $\dim(\text{Span}\{B_{\pi(S_i)}\}) = k$ for all $i \in [m]$. Fix $i \in [m]$ and note that by (12) we have for all unit vectors $\mathbf{u} \in \text{Span}\{A_{S_i}\}$ that $d(u, \text{Span}\{B_{\pi(S_i)}\}) \leq \frac{\phi_k(A)}{\rho k}\delta$ for $\delta < \frac{L_2(A)}{\sqrt{2}}$. By definition of $L_2(A)$ we have for all 2-sparse $\mathbf{x} \in \mathbb{R}^m$:

$$L_2(A) \leq \frac{|A\mathbf{x}|_2}{|\mathbf{x}|_2} \leq \rho\frac{|\mathbf{x}|_1}{|\mathbf{x}|_2} \leq \rho\sqrt{2} \qquad (36)$$

Hence $\delta < \rho$. Since $\phi_k \leq 1$ we have $d(u, \text{Span}\{B_{\pi(S_i)}\}) < 1$ and it follows by Lemma 2 that $\dim(\text{Span}\{B_{\pi(S_i)}\}) \geq \dim(\text{Span}\{A_{S_i}\}) = k$. Since $|\pi(S_i)| = k$, we in fact have $\dim(\text{Span}\{B_{\pi(S_i)}\}) = k$.

We will now show that

$$|\bigcap_{i\in J} \pi(S_i)| \leq |\bigcap_{i\in J} S_i| \quad \text{for all } J \in \binom{[m]}{k}. \qquad (37)$$

Fix $J \in \binom{[m]}{k}$. By (12) we have for all unit vectors $\mathbf{u} \in \cap_{i\in J}\text{Span}\{B_{\pi(S_i)}\}$ that $d(\mathbf{u}, \text{Span}\{A_{S_i}\}) \leq \frac{\phi_k(A)}{\rho k}\delta$ for all $j \in J$, where $\delta < \frac{L_2(A)}{\sqrt{2}}$. It follows by Lemma 6 that

$$d\left(\mathbf{u}, \bigcap_{i\in J} \text{Span}\{A_{S_j}\}\right) \leq \frac{\delta}{\rho}\left(\frac{\phi_k(A)}{1 - \xi(\{\text{Span}\{A_{S_i}\} : i \in J\})}\right) \leq \frac{\delta}{\rho},$$

where the second inequality follows immediately from the definition of $\phi_k(A)$.

Now, since $\text{Span}\{B_{\cap_{i\in J}\pi(S_i)}\} \subseteq \cap_{i\in J}\text{Span}\{B_{\pi(S_i)}\}$ and (by Lemma 5) $\cap_{i\in J}\text{Span}\{A_{S_i}\} = \text{Span}\{A_{\cap_{i\in J}S_i}\}$, we have

$$d\left(\mathbf{u}, \text{Span}\{A_{\cap_{i\in J}S_i}\}\right) \leq \frac{\delta}{\rho} \quad \text{for all unit vectors } \mathbf{u} \in \text{Span}\{B_{\cap_{i\in J}\pi(S_i)}\}. \qquad (38)$$

We therefore have by Lemma 2 (since $\delta/\rho < 1$) that $\dim(\text{Span}\{B_{\cap_{i\in J}\pi(S_i)}\}) \leq \dim(\text{Span}\{A_{\cap_{i\in J}S_i}\})$ and (37) follows by the linear independence of the columns of A_{S_i} and $B_{\pi(S_i)}$ for all $i \in [m]$.

Suppose now that $J = \{i-k+1, \ldots, i\}$ so that $\cap_{i\in J}S_i = i$. By (37) we have that $\cap_{i\in J}\pi(S_i)$ is either empty or it contains a single element. Lemma 7 ensures that the latter case is the only possibility. Thus, the association $i \mapsto \cap_{i\in J}\pi(S_i)$ defines a map $\hat{\pi} : [m] \to [m]$. Recalling (14), it follows from (38) that for all unit vectors $\mathbf{u} \in \text{Span}\{A_i\}$ we have $d\left(\mathbf{u}, \text{Span}\{B_{\hat{\pi}(i)}\}\right) \leq \delta/\rho$ also. Since i is arbitrary, it follows that for every canonical basis vector $\mathbf{e}_i \in \mathbb{R}^m$, letting $c_i = |A\mathbf{e}_i|_2^{-1}$ and $\varepsilon = \delta/\rho$, there exists some $c_i' \in \mathbb{R}$ such that $|c_iA\mathbf{e}_i - c_i'B\mathbf{e}_{\hat{\pi}(i)}|_2 \leq \varepsilon$ where $\varepsilon < \frac{L_2(A)}{\sqrt{2}} \min_{j\in[m]} c_i$. This is exactly the supposition in (17) and the result follows from the subsequent arguments of Section III. □

Remark 6. *The arguments above can easily be modified to prove Lemma 3. Since Lemma 7 assumes $m = m'$, we may not invoke it when $m' > m$ to show that $|\cap_{i\in J} \pi(S_i)| = 1$ for $J = \{i - k + 1, \ldots, i\}$. Instead, under the additional assumption that B satisfies spark condition* (**??**)*, we can swap the roles of A and B in the proof of* (38) *to show that $\dim(\text{Span}\{B_{\cap_{i\in J}\pi(S_i)}\}) = \dim(\text{Span}\{A_{\cap_{i\in J}S_i}\})$, from which the same fact then follows. The proof is then completed in much the same way as above by defining a map $\pi : [m] \to [m']$ by the association $i \mapsto \cap_{i\in J}\pi(S_i)$, thereby reducing the proof to the $k = 1$ case described in Remark 3.*

Remark 7. *In general, there may exist combinations of fewer supports with intersection $\{i\}$, e.g. if $m \geq 2k - 1$ then $S_{i-(k-1)} \cap S_i = \{i\}$. For brevity, we have considered a construction that is valid for any $k < m$.*

Acknowledgment

We thank Fritz Sommer for turning our attention to the dictionary learning problem, Darren Rhea for sharing his early explorations of this problem, and Ian Morris for posting identity (14) with a reference to his proof of it on the internet (www.stackexchange.com).

Charles J. Garfinkle completed a B.S. in Physics and Chemistry at McGill University. He is currently a Ph.D. candidate in Neuroscience at UC Berkeley.

Christopher J. Hillar completed a B.S. in Mathematics and a B.S. in Computer Science at Yale University. Supported by an NSF Graduate Research Fellowship, he received his Ph.D. in Mathematics from the University of California (UC), Berkeley in 2005. From 2005-2008, he was a Visiting Assistant Professor and NSF Postdoctoral Fellow at Texas A&M University. From 2008-2010, he was an NSF Mathematical Sciences Research Institutes Postdoctoral Fellow at the Mathematical Sciences Research Institute (MSRI) in Berkeley, CA. In 2010, he joined the Redwood Center for Theoretical Neuroscience at UC Berkeley, and in 2011, he became a research specialist in the Tecott mouse behavioral neuroscience lab at UC San Francisco.