Title
A Tale of Two Dictionary Learning Problems

Author
Garfinkle, Charles Joseph

Publication Date
2020

A Tale of Two Dictionary Learning Problems

by

Charles J. 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:

Adjunct Professor Friedrich Sommer, Chair
Professor Bruno Olshausen
Associate Professor Michael DeWeese
Chancellor's Professor Bin Yu

Summer 2020

A Tale of Two Dictionary Learning Problems

Abstract

A Tale of Two Dictionary Learning Problems

by

Charles J. Garfinkle

Doctor of Philosophy in Neuroscience

University of California, Berkeley

Adjunct Professor Friedrich Sommer, Chair

Learning optimal dictionaries for sparse representation modeling has led to the discovery of characteristic sparse features in several classes of natural signals. However, universal guarantees of the uniqueness and stability of such features in the presence of noise are lacking. This work presents very general conditions guaranteeing when dictionaries yielding the sparsest encodings of a dataset are unique and stable with respect to noise. The stability constants are explicit and computable; as such, there is an effective procedure sufficient to affirm if a proposed solution to the dictionary learning problem is unique within bounds commensurate with the noise.

Two formulations of the dictionary learning problem are considered. The first seeks a dictionary for which each signal in a dataset is approximated up to some bounded error by a linear superposition of only a limited number of dictionary elements. In this case, existing guarantees are extended to the noisy regime to show that such dictionaries and the sparse representations they induce are almost always identifiable up to an error commensurate with the approximation error. Moreover, a theory of combinatorial designs is introduced to demonstrate that this is so even if the dictionary fails to satisfy the spark condition, the data are distributed over only a polynomial set of subspaces spanned by the dictionary, or (to some extent) even if the dictionary is overestimated in size.

The second formulation of the problem seeks a dictionary which minimizes the average number of dictionary elements required to approximate each signal in the dataset up to some bounded error. The guarantees in this case, the first of their kind in both in the noiseless and noisy regimes, are derived by demonstrating that this second problem actually reduces to an instance of the first. Importantly, in both cases, no constraints whatsoever are imposed on learned dictionaries beyond a natural upper bound on their size.

This work serves to justify, in principle, dictionary learning in general as a means of discovering latent sparse structure in real data. Though much work remains to be done deriving criteria for use in practice, the theoretical tools developed here should be of use to this end.

I want to dedicate this work to my dad. Yet it feels as though doing so would in some sense measure his life up against these results, which can't possibly be matched to the time with him I traded to produce them. Whatever the eventual effect of this work on the scientific ecosystem may be, these mathematical truths could have waited. They could have arrived at their own pace, via myself or another, and I could have been more present with my dad as he waited patiently to die of cancer.

I suppose that I couldn't really fathom what it actually meant for this to happen – to never see or talk to or hug him again – while the idea of putting original work out there, of crafting my own piece of the overwhelming puzzle that is neuroscience, was tantalizing. Yet I feel that it was very much a sense of desperation, of having to accomplish something and prove myself worthy of the position I had come to find myself in after 20+ years of schooling and a first failed project, that largely drove these efforts. Truth be told, I initially had no particular interest in dictionary learning other than as a means to passing my qualifying exam; and by that time, I worried the brain was so dauntingly complex that I should feel lucky to have at least been offered a well-defined problem of arguable relevance to neuroscience to work on, particularly one I felt even remotely equipped to solve myself.

So let's just say that this is the first bit of a life's work dedicated to my dad. A life of work uninfluenced by position or status and the associated self-doubt, pressure, or delusions of grandeur; one dedicated to passionate ideas, to solving problems I care about and that really matter, to loving and caring for others or, as my dad would say, to "being a mensch". May this dedication serve to hold me accountable for that. Also, I love you, mom. I dedicate this and everything to come to you, too.

Contents

List of Figures

Acknowledgments

First and foremost, I would like to thank Chris Hillar for clearing the path in his original work on the dictionary problem, and for vetting all of my proofs many, many times over. He asked for nothing in return for his time, attention, and mathematical expertise, and continues to extend this generosity to anyone nerdy enough to handle it.

Thank you to Fritz Sommer for recommending my acceptance to Berkeley and for advising me as his student; although I eventually had to leave his lab to work on other problems of interest to me, he kindly lent me his time over the course of four years. To the other members of my thesis committee: Bruno Olshausen, Mike DeWeese, and Bin Yu, for their time and consideration over the years as well, and for approving this dissertation. My mom would have killed me if I didn't claim this degree, so they saved my life. To the graduate program manager of the Helen Wills Neuroscience Institute, Candace Groskreutz, for her kindness, patience, and support. To the HWNI itself, the Redwood Center for Theoretical Neuroscience, and the friends I made in the lab and elsewhere; I hope we remain in each other's lives.

Finally, a heartfelt thank you to Eric Weiss in particular for introducing me to the student co-ops, and to the Berkeley Student Cooperative (www.bsc.coop) for the work they do providing the amazing experience of living co-operatively in Berkeley. The co-ops changed my life, and I am forever grateful – thank you so, so much to the hundreds of talented, creative, weird, passionate, intelligent, kind, thoughtful, absurd, and inspiring people I have lived with throughout the years for all of the wonderful and weird times. CZ or die.

Chapter 1

Introduction

It is a long-standing practice in the field of signal processing to describe signals as linear combinations of elementary waveforms selected from a pre-specified “dictionary”. Every signal has a unique representation in terms of these components when the dictionary forms a basis. For example, a signal can be decomposed into its constituent frequencies via the Fourier transform, which performs a change of basis.

Bases have been the dominant form of signal representation until recently due largely to their simplicity. For many signal analysis tasks, however, no one basis can convey clearly all of the relevant information in the signal. For instance, if a signal can be either a sine wave or a delta function, then neither the standard basis nor the Fourier basis will indicate one case as explicitly as it does the other.

This need for more freedom of expression has led to the development of redundant signal representations utilizing overcomplete dictionaries, which contain more waveforms than there are dimensions of the signal. An overcomplete dictionary admits infinitely many possible ways to decompose a signal into a linear superposition its constituent components. The intention is then to seek the best such representation, be it by some analytic criteria, or as measured by some task-specific cost function.

A popular approach to the design of overcomplete dictionaries in the latter case has been to seek one which admits a sparse representation for every signal of interest; that is, each signal can be reconstructed, or at least well-approximated, by a combination of only a few dictionary elements from the bunch. Carrying on with our running example, the union of the standard basis with the Fourier basis is an overcomplete dictionary with respect to which both sine waves and delta functions achieve the sparsest possible representation as the scaling of a single elementary waveform from the dictionary.

Early approaches to sparse representation modeling would assume a model of the signal class from which a suitable sparsifying dictionary could be derived, as we have in our trivial example. Such dictionaries are characterized by an analytic formulation and a fast implicit implementation; yet they tend to be over-simplistic when applied to model natural phenomena.

An alternative modern approach to dictionary design is conditioned on the assumption that the sparse structure of signals conveying information about complex natural phenomena can be more accurately extracted directly from a training dataset, a process referred to as dictionary learning (see [50] for a comprehensive review). In the seminal work [38] (see also [26, 7, 22]), a dictionary trained over a collection of small patches extracted from images of the natural environment was shown to share qualitative similarities with linear filters estimated from the response properties of simple-cell neurons in mammalian visual cortex, which until then had been more weakly described analytically as Gabor filters. This remarkable discovery demonstrated that the assumption of sparsity alone could potentially account for a fundamental property of the visual system, and showcased the potential of the machine-learning approach to dictionary design. Even more curiously, these waveforms (e.g., Gabor-like wavelets) tend to appear in dictionaries optimized with respect to different natural image datasets by a variety of dictionary learning algorithms, suggesting that the optimal dictionaries for sparse representation of these signals may, in some sense, be canonical [11].

In light of these observations, it is natural to wonder when the optimal dictionary for sparse representation modeling can be identified given a representative sample from a signal class. Answers to this question have implications in practice wherever an appeal is made to latent sparse structure of data (e.g., forgery detection [25, 39]; brain recordings [30, 1, 33]; and gene expression [49]), since the assumption is that this structure captures some identifiable physical or logically causal variable.

Even though several dictionary learning algorithms have recently been proposed to provably recover a unique dictionary under specific conditions (see [45, Sec. I-E] for a summary of the state of the art), few theorems can be invoked to justify inference with respect to this model of data more broadly. Despite the now ubiquitous application of dictionary learning methods in practice, to the best of my knowledge a universal guarantee of the uniqueness and stability of learned dictionaries and the sparse representations they induce over real data in the presence of noise has yet to appear in the literature.

In this work, it is proven very generally that uniqueness and stability is a typical property of learned dictionaries. Specifically, if each of N observed n-dimensional real signals is truly a (noisy) linear combination of at most k elementary waveforms drawn from a suitable dictionary of size $m \ll N$, that dictionary is uniquely specified by the data up to an error that is linear in the noise given $N = m(k-1)\binom{m}{k} + m$ such signals (Thm. 1 and Cor. 1). In fact, provided $n \geq \min(2k, m)$, in almost all cases the problem is well-posed, as per Hadamard [21], given enough data (Thm. 3 and Cor. 2). Similar guarantees also hold for the related (and perhaps more commonly posed, e.g. [42]) optimization problem seeking a dictionary minimizing the average number of elementary waveforms required to reconstruct each sample of the dataset (Thm. 2). To great practical benefit (and technical pain!), these guarantees apply without the imposition of any constraints at all on learned dictionaries beyond an upper bound on their size, which is necessary in any case to avoid trivial solutions (e.g., allowing $m = N$).

1.1 The dictionary learning problem(s)

Let us now rigorously define the two formulations of the dictionary learning problem with which this work concerns itself. Fix a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ with the elementary waveforms of the dictionary as its columns $\mathbf{A}_j$ ($j = 1, \ldots, m$) and let dataset Z consist of measurements:

$$\mathbf{z}_i = \mathbf{A}\mathbf{x}_i + \mathbf{n}_i, \quad i = 1, \ldots, N, \tag{1.1}$$

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 shall first consider the following decidable[1] formulation of the dictionary learning problem.

Problem 1. *Find a 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., [35, 18, 2, 24]) 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*:

$$\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_2 \implies \mathbf{x}_1 = \mathbf{x}_2, \quad \text{for all } k\text{-sparse } \mathbf{x}_1, \mathbf{x}_2, \tag{1.2}$$

which is necessary to guarantee the uniqueness of arbitrary k-sparse $\mathbf{x}_i$. We shall 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:*

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

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

$$\|\mathbf{A}_j - (\mathbf{BPD})_j\|_2 \leq \delta_1 \text{ and } \|\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_i\|_1 \leq \delta_2. \tag{1.3}$$

[1]Note that Prob. 1 is decidable for rational inputs $\mathbf{z}_i$ [23] since the statement that it has a solution can be expressed as a logical sentence in the theory of algebraically closed fields, and this theory has quantifier elimination [5].

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 (1.3). Consider any dataset Z generated as in (1.1) 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.[2]

The main result of this work is a very general uniqueness theorem (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. 2.1). Moreover, an explicit, computable $\varepsilon(\delta_1, \delta_2)$ is given in (2.4) that is linear in desired accuracy δ_1, and essentially so in δ_2.

The same guarantees can be extended (Thm. 2) to the following alternate formulation of the dictionary learning problem, which seeks to minimize the average number of nonzero entries in sparse codes.

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

$$\min \sum_{i=1}^{N} \|\overline{\mathbf{x}}_i\|_0 \quad \textit{subject to} \quad \|\mathbf{z}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \eta, \ \textit{for all } i. \tag{1.4}$$

Surprisingly, the development of Thm. 1 is general enough to guarantee uniqueness and stability even when generating $\mathbf{A}$ do not fully satisfy (1.2), and to some degree even when recovery dictionaries $\mathbf{B}$ have more columns than $\mathbf{A}$. The approach incorporates a theory of combinatorial designs for the sparse supports of generating codes $\mathbf{x}_i$ that should also be of independent interest.

1.2 Outline of the thesis

Formal statements of the main findings described above are given in Chap. 2, along with their adaptation to dictionaries and codes drawn from arbitrary (continuous) probability distributions (Thm. 3 and Cor. 2). All results assume real matrices and vectors. For clarity of exposition, the technical proofs of Thms. 1 and 2 are deferred to Chap. 3, following some necessary definitions and the statement of a key lemma in combinatorial matrix analysis (Lem. 1, proven in the chapter's Appendix). These results and their applications are discussed in Chap. 4, which concludes with some open questions and directions for future research, which are seeded in part by some practically-minded simulations.

[2]We mention that the different norms in (1.3) reflect the distinct meanings typically ascribed to the dictionary and sparse codes in modeling data.

Chapter 2

Results

2.1 Definitions

To state precisely the results of this work, we will require first the identification of some combinatorial criteria on the supports[1] of sparse vectors. Let $\{1, \ldots, m\}$ 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 $\mathcal{H} \subseteq 2^{[m]}$. Let us 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*). Let us 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 $i \in [m]$.*

Next, we will require a quantitative generalization of the spark condition (1.2) to combinatorial subsets of supports. The *lower bound* of an $n \times m$ matrix $\mathbf{M}$ is the largest α with $\|\mathbf{Mx}\|_2 \geq \alpha\|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{R}^m$ [20]. By compactness of the unit sphere, every injective linear map has a positive lower bound; hence, if $\mathbf{M}$ satisfies (1.2), 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 (2.1) by restricting it to the spans of certain submatrices[2] 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, let $\boldsymbol{\mathcal{M}}_S$ denote the column-span of a submatrix $\mathbf{M}_S$, and $\boldsymbol{\mathcal{M}}_\mathcal{G}$ to denote $\{\boldsymbol{\mathcal{M}}_S\}_{S \in \mathcal{G}}$. Define:

$$L_{\mathcal{H}}(\mathbf{M}) := \min\left\{\frac{\|\mathbf{M}_S\mathbf{x}\|_2}{\sqrt{k}\|\mathbf{x}\|_2} : S \in \mathcal{H},\ \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^{|S|}\right\}, \tag{2.1}$$

[1] 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.

[2] See [48] for an overview of the related "union of subspaces" model.

writing also L_k in place of $L_{\mathcal{H}}$ when $\mathcal{H} = \binom{[m]}{k}$.[3] As explained above, compactness implies that $L_{2k}(\mathbf{M}) > 0$ for all $\mathbf{M}$ satisfying (1.2). 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 $2k > m$).

2.2 Uniqueness theorems

Deterministic guarantees

The following is the statement of the main result. For expository purposes the quantity C_1 (a function of $\mathbf{A}$, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, and $\mathcal{H}$) will be left undefined until Chap. 3.

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 $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ include more than $(k-1)\binom{\overline{m}}{k}$ vectors in general linear position[4] supported in each $S \in \mathcal{H}$, then the following recovery guarantees hold for $C_1 > 0$ given by* (3.11).

Dictionary Recovery: *Fix $\varepsilon < L_2(\mathbf{A})/C_1$.[5] 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:*

$$\|\mathbf{A}_j - (\mathbf{BPD})_j\|_2 \leq C_1\varepsilon, \quad \textit{for all } j \in J, \tag{2.2}$$

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

Code Recovery: *If, moreover, $\mathbf{A}_J$ satisfies* (1.2) *and $\varepsilon < L_{2k}(\mathbf{A}_J)/C_1$, then $(\mathbf{BP})_J$ also satisfies* (1.2) *with $L_{2k}(\mathbf{BP}_J) \geq (L_{2k}(\mathbf{A}_J) - C_1\varepsilon)/\|\mathbf{D}_J\|_1$, and for all $i \in [N]$:*

$$\|(\mathbf{x}_i)_J - (\mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_i)_J\|_1 \leq \left(\frac{1 + C_1\|(\mathbf{x}_i)_J\|_1}{L_{2k}(\mathbf{A}_J) - C_1\varepsilon}\right)\varepsilon, \tag{2.3}$$

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

[3] 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 [8].

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

[5] 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 (2.2).

$\mathbf{x}_i$ 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 (1.3) guaranteed in Def. 1 for the following worst-case error ε:

$$\varepsilon(\delta_1, \delta_2) := \min\left\{\frac{\delta_1}{C_1}, \frac{\delta_2 L_{2k}(\mathbf{A})}{1 + C_1\left(\delta_2 + \max_{i\in[N]} \|\mathbf{x}_i\|_1\right)}\right\}. \tag{2.4}$$

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* (1.2) *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* (2.4).

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.

It should be stressed here that framing the problem in terms of hypergraphs will allowed us to show, unlike in previous research on the subject, that the matrix $\mathbf{A}$ need not necessarily satisfy (1.2) 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 (1.2) 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. A brief discussion on this point is deferred until later.

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 the dictionary recovery and code recovery guarantees of Thm. 1 when the corresponding bounds on ε are met.*

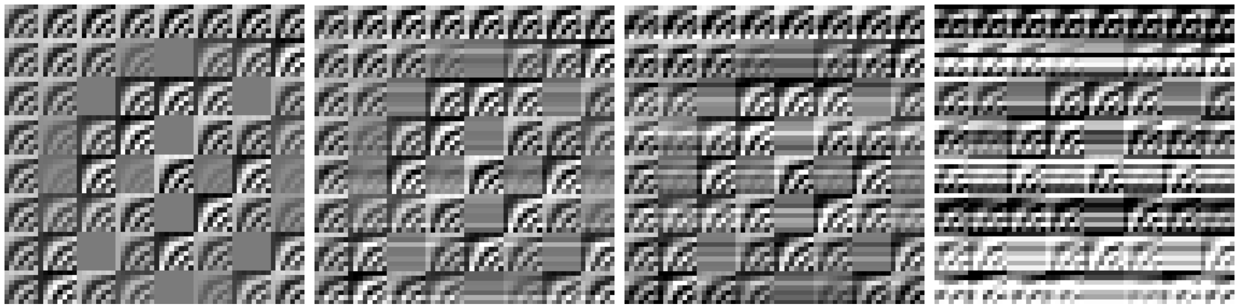

Figure 2.1: **Learning a dictionary from increasingly noisy data**. The (unraveled) basis elements of the 8×8 discrete cosine transform (DCT) form the 64 columns of the left-most matrix above. Three increasingly imprecise dictionaries (columns reordered to best match original) are recovered by FastICA [28] trained on data generated from 8-sparse linear combinations of DCT elements corrupted with additive noise (increasing from left to right).

Probabilistic guarantees

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 (1.2) 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 $(i, j) \in S' \times S$.[6]

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 (1.2).

It turns out that a similar phenomenon applies to datasets of vectors with a stable sparse representation. Briefly, following the same procedure as in [24, 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

[6]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 [47].

$n \times N$ 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 [15]). This fact combined with Thm. 3 implies the following.[7]

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.*

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

2.3 Discussion

It is befitting to comment on the optimality of these results. The linear scaling for ε in (2.4) is essentially optimal (e.g., see [3]), but a basic open problem remains: how many samples are necessary to determine the sparse coding model? These 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 my 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.[8] Reasons to be skeptical that this holds in general, however, can be found in [47, 46]. Even so, the next chapter contains a discussion on how probabilistic guarantees can in fact be made for any number of available samples (see also Fig. 4.1).

As it seemed to benefit a reviewer of this work, some clarification may be in order on how the deterministic sample complexity $N = |\mathcal{H}|\left[(k-1)\binom{m}{k}+1\right]$ given here (Cor. 1) compares to those listed in the the top two rows listed in Table I of [24]. To be clear, the theory

[7]We refer the reader to [24] for a more detailed explanation of these arguments.

[8]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.

developed here is strictly more general, since $\mathcal{H}$ can always be taken to be $\binom{[m]}{k}$. The point of difference in this comparison is the assumed set of supports for sparse codes, which is always $\binom{[m]}{k}$ in [24], whereas here it can be assumed to be any regular k-uniform hypergraph that satisfies the SIP. By row,

I. The result here improves upon the listed $k\binom{m}{k}^2$ by an exponential factor, since for every $k < m$ there exists a regular k-uniform hypergraph $\mathcal{H}$ with $|\mathcal{H}| = m$ satisfying the SIP.

II. The authors in [24] have applied measure-theoretic arguments to achieve $(k+1)\binom{m}{k}$ with almost-certainty (i.e. with probability one), a factor of m reduction over that for which certainty can alternatively be guaranteed here. In their case, $(k+1)$ vectors need to be allocated to each of $\binom{m}{k}$ distinct supports, whereas here $(k-1)\binom{m}{k}+1$ vectors are allocated to each of m distinct supports.

Chapter 3

Proofs

We shall prove Thm. 1 and then prove Thm. 2 by arguing that Prob. 2 reduces to an instance of Prob. 1 given sufficient data.

It is instructive to begin the proof of Thm. 1 by showing how dictionary recovery (2.2) already implies sparse code recovery (2.3) when $\mathbf{A}$ satisfies (1.2) and $\varepsilon < L_{2k}(\mathbf{A})/C_1$. We shall 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 (2.1), and noting that $\sqrt{k}\|\mathbf{v}\|_2 \geq \|\mathbf{v}\|_1$ for k-sparse $\mathbf{v}$, we have for all $i \in [N[$:

$$\begin{aligned}
\|\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_i\|_1 &\leq \frac{\|\mathbf{BPD}(\mathbf{x}_i - \mathbf{D}^{-1}\mathbf{P}^\top\overline{\mathbf{x}}_i)\|_2}{L_{2k}(\mathbf{BPD})} \\
&\leq \frac{\|(\mathbf{BPD} - \mathbf{A})\mathbf{x}_i\|_2 + \|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2}{L_{2k}(\mathbf{BPD})} \\
&\leq \frac{C_1\varepsilon\|\mathbf{x}_i\|_1 + \varepsilon}{L_{2k}(\mathbf{BPD})},
\end{aligned} \tag{3.1}$$

where the first term in the numerator above follows from the triangle inequality and (2.2).

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

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

We therefore have that $L_{2k}(\mathbf{BPD}) \geq L_{2k}(\mathbf{A}) - C_1\varepsilon > 0$, and (2.3) then follows from (3.1). 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 (2.2), which we shall now establish beginning with the important special case of $k = 1$.

3.1 Proving the case $k = 1$

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 $\varepsilon < L_2(\mathbf{A}) \min_{\ell\in[m]} |c_\ell|$.

Proof of Thm. 1 for $k = 1$. 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:

$$\|c_i\mathbf{A}_i - \overline{c}_i\mathbf{B}_{\pi(i)}\|_2 \leq \varepsilon, \quad \text{for } i \in [m]. \tag{3.2}$$

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

Let us 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:

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

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 (3.2) to see that for all $i \in [m]$:

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

$\square$

3.2 Stating the main lemma

An extension of the proof to the general case $k < m$ requires some additional tools to derive the general expression (3.11) 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), which contains most of the complexity of the proof of Thm. 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:*

$$d(U, V) := \max_{\mathbf{u}\in U,\ \|\mathbf{u}\|_2\leq 1} dist(\mathbf{u}, V). \tag{3.3}$$

Note the following facts about d. Clearly,

$$U' \subseteq U \implies d(U', V) \leq d(U, V). \tag{3.4}$$

From [31, Ch. 4 Cor. 2.6], we also have:

$$d(U, V) < 1 \implies \dim(U) \leq \dim(V), \tag{3.5}$$

and from [37, Lem. 3.2]:

$$\dim(U) = \dim(V) \implies d(U, V) = d(V, U). \tag{3.6}$$

The required stability result in combinatorial matrix analysis is the following. For expository purposes, the proof of this fact is relegated to Sec. 3.5.

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* (3.10)*:*

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

$$d(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\pi(S)}) \leq \varepsilon, \quad \textit{for } S \in \mathcal{H}, \tag{3.7}$$

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:

$$\|\mathbf{A}_i - (\mathbf{BPD})_i\|_2 \leq C_2\varepsilon, \quad \textit{for } i \in J, \tag{3.8}$$

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

The constant C_2 (a function of $\mathbf{A}$ and $\mathcal{H}$) will be presented relative to a quantity used in [10] to analyze the convergence of the "alternating projections" algorithm for projecting a point onto the intersection of subspaces. This quantity is incorporated into the following definition, which we shall refer to in the proof of Lem. 3 in the Sec. 3.5; specifically, it will be used 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:*

$$\xi^2(\mathcal{V}) := 1 - \max \prod_{i=1}^{\ell-1} \sin^2 \theta\left(V_i, \cap_{j>i} V_j\right), \tag{3.9}$$

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 [10, Def. 9.4]:

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

Note that $\theta \in (0, \frac{\pi}{2}]$ implies $0 \le \xi < 1$, and that $\xi(\mathcal{V}') \le \xi(\mathcal{V})$ when $\mathcal{V}' \subseteq \mathcal{V}$.[1]
The constant $C_2 > 0$ of Lem. 1 can now be expressed as:

$$C_2(\mathbf{A}, \mathcal{H}) := \frac{(r+1)\max_{j\in[m]} \|\mathbf{A}_j\|_2}{1 - \max_{\mathcal{G}\in\binom{\mathcal{H}}{r+1}} \xi(\boldsymbol{\mathcal{A}}_\mathcal{G})}. \tag{3.10}$$

3.3 Proving the general case $k < m$

We may now 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. Define:

$$C_1(\mathbf{A}, \mathcal{H}, \{\mathbf{x}_i\}_{i=1}^N) := \frac{C_2(\mathbf{A}, \mathcal{H})}{\min_{S\in\mathcal{H}} L_k(\mathbf{A}\mathbf{X}_{I(S)})}. \tag{3.11}$$

Given the assumptions of Thm. 1 on $\mathbf{A}$ and the $\mathbf{x}_i$, this expression for C_1 is well-defined[2] and yields an upper bound on ε consistent with that proven sufficient in the case $k = 1$ considered at the beginning of this chapter.[3]

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 [47], although efficiently computable bounds have been proposed (e.g., via the "mutual coherence" of a matrix [12]); 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.

Proof of Thm. 1 for $k < m$. We shall 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 $(k-1)\binom{m}{k}$ 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[4] from $L_{2\mathcal{H}}(\mathbf{A}) > 0$ and the general

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

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

[3] When $\mathbf{x}_i = c_i \mathbf{e}_i$, we have $C_2 \ge 2\|\mathbf{A}_i\|_2$ and the denominator in (3.11) becomes $\min_{i\in[m]} |c_i|\|\mathbf{A}_i\|_2$; hence, $C_1 \ge 2/\min_{i\in[m]} |c_i|$.

[4] See footnote 2.

linear position of the $\mathbf{x}_i$ that $L_k(\mathbf{AX}_K) > 0$; that is, the columns of the $n \times k$ matrix $\mathbf{AX}_K$ 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 $\mathbf{y} = \mathbf{AX}_K\mathbf{c}$. 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:

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

where the last inequality uses (2.1). From Def. 3:

$$d(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\overline{S}}) \leq \frac{\varepsilon}{L_k(\mathbf{AX}_K)} \leq \frac{\varepsilon}{L_k(\mathbf{AX}_{I(S)})} \leq \varepsilon\frac{C_1}{C_2}. \tag{3.12}$$

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, let us briefly revisit the discussion on sample complexity from the end of the previous chapter. 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.[5] 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 (see Fig. 4.1).

Proof of Thm. 2. We shall 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 (1.4) we have:

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

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:

$$\sum_{p=k+1}^{\overline{m}} n_p \leq \sum_{p=k+1}^{\overline{m}} (p-k)n_p \leq \sum_{p=0}^{k}(k-p)n_p \leq k\sum_{p=0}^{k-1} n_p, \tag{3.13}$$

[5] 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%.

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, (3.12) 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 (3.12) is less than one for $\varepsilon < L_2(\mathbf{A})/C_1$. Thus, by (3.5) 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 (3.13), 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 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 $i = 1, \ldots, N$. □

3.4 Discussion

The absence of any assumptions at all about dictionaries that solve Prob. 1 was a major technical hurdle in proving Thm. 1. This very general guarantee was sought because of the practical difficulty of ensuring that an algorithm maintain a dictionary satisfying the spark condition (1.2) at each iteration, which (to my knowledge) has been an explicit or implicit assumption of all previous works except [24]; indeed, even certifying a dictionary has this property is NP-hard [47].

Several results in the literature had to be combined to extend the guarantees derived in [24] into the noisy regime. The main challenge was to generalize Lem. 1 to the case where the k-dimensional subspaces spanned by corresponding submatrices $\mathbf{A}_S$ and $\mathbf{B}_{\pi(S)}$ are assumed to be"close" but not identical. Referring now to the proof in this chapter's Appendix, the situation is unlike that in [24], where an inductive argument could be applied to the noiseless case. Rather, here it has to be explicitly demonstrated that this proximity relation is propagated through iterated intersections right down to the spans of the dictionary elements themselves. Lem. 3 was designed to encapsulate this fact, proven by appeal to a convergence guarantee for an alternating projections algorithm first proposed by von Neumann. This result, combined with a little known fact (3.6) about the distance metric between subspaces, make up the more obscure components of the deduction.

The proof of Lem. 1 diverges most significantly from the approach taken in [24] by way of Lem. 4, which utilizes a combinatorial design for support sets (the "singleton intersection property") to reduce the deterministic sample complexity by an exponential factor. This constitutes a significant advance toward legitimizing dictionary learning in practice, since data must otherwise exhibit the exponentially many possible k-wise combinations of dictionary elements required by (to my knowledge) all previously published results; although an exponential number of samples per support is still required (unless k is held fixed). The issue is that the map $\pi : \mathcal{H} \to \binom{[m]}{k}$ is surjective only when $\mathcal{H}$ is taken to be $\binom{[m]}{k}$, in which case one may proceed by induction as in [24], freely choosing supports in the codomain of π to intersect over $(k-1)$ indices to then map back to some corresponding set of $(k-1)$ indices at the intersection of supports in the domain. Here, for $\mathcal{H} \subset \binom{[m]}{k}$, a bijection between indices had to instead be established by pigeonholing the image of π under constraints imposed by the SIP, which was formulated specifically for this purpose. It just so happened that this more general argument for a non-surjective π constrained by the SIP applied just as well to the situation where the number of dictionary elements m is over-estimated (i.e. $\overline{m} > m$), in which case a one-to-one correspondence can be guaranteed between a subset of columns of $\mathbf{A}$ and $\mathbf{B}$ of a size simply expressed in terms of the width of each matrix and the regularity of $\mathcal{H}$.

One of the mathematically significant achievements in [24] was to break free from the constraint that the recovery matrix $\mathbf{B}$ satisfy the spark condition in addition to the generating dictionary $\mathbf{A}$. Here, it has been demonstrated that, in fact, even $\mathbf{A}$ need not satisfy the spark condition! Rather, $\mathbf{A}$ need only be injective on the union of subspaces with supports forming a regular k-uniform hypergraph satisfying the SIP (a distinguishing example is given in Sec. 2). This relaxation of constraints inspired the definition of the restricted matrix lower bound $L_{\mathcal{H}}$ in (2.1), which generalizes the well-known (see footnote 3) restricted matrix lower bound L_k to be in terms of a hypergraph $\mathcal{H}$, and is an interesting object for further study in its own right.

To reiterate, the methods applied here to prove Thm. 1 yield the following results beyond a straightforward extension of those in [24] to the noisy case:

1. A reduction in deterministic sample complexity: To identify the $n \times m$ generating dictionary $\mathbf{A}$, it is required in [24] that $k\binom{m}{k}$ data points be sampled from each of the $\binom{m}{k}$ subspaces spanned by subsets of k columns of $\mathbf{A}$. It is shown here that in fact it suffices to sample from at most m such subspaces (see Cor. 1).

2. An extension of guarantees to the case where the number of dictionary elements is unknown: The results of [24] only apply to the case where the matrix $\mathbf{B}$ has the same number of columns as $\mathbf{A}$. It is shown here that if $\mathbf{B}$ has at least as many columns as $\mathbf{A}$ then it contains (up to noise) a subset of the columns of $\mathbf{A}$.

3. Relaxed requirements (no spark condition) on the generating matrix $\mathbf{A}$: Rather, $\mathbf{A}$ need only be injective on the union of subspaces with supports that form a regular uniform hypergraph satisfying the SIP.

3.5 Appendix: Proving the main lemma

We shall prove Lem. 1 after the following auxiliary lemmas.

Lemma 2. *If $f : V \to W$ is injective, then $f\left(\cap_{i=1}^{\ell} V_i\right) = \cap_{i=1}^{\ell} f(V_i)$ 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 $w \in f(U \cap V)$. $\square$

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):*

$$dist(\mathbf{u}, V) \leq \frac{1}{1 - \xi(\mathcal{V})} \sum_{i=1}^{k} dist(\mathbf{u}, V_i). \tag{3.14}$$

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:

$$\begin{aligned}\|\mathbf{u}-\Pi_V\mathbf{u}\|_2 &\leq \|\mathbf{u} - \Pi_{V_k}\mathbf{u}\|_2 + \|\Pi_{V_k}\mathbf{u} - \Pi_{V_k}\Pi_{V_{k-1}}\mathbf{u}\|_2 \\ &\quad + \cdots + \|\Pi_{V_k}\Pi_{V_{k-1}} \cdots \Pi_{V_1}\mathbf{u} - \Pi_V\mathbf{u}\|_2 \\ &\leq \sum_{\ell=1}^{k} \|\mathbf{u} - \Pi_{V_\ell}\mathbf{u}\|_2 + \|(\Pi_{V_k} \cdots \Pi_{V_1} - \Pi_V)\mathbf{u}\|_2, \end{aligned} \tag{3.15}$$

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 (3.15) 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 (3.14) follows from [10, Thm. 9.33]:

$$\|\Pi_{V_k}\Pi_{V_{k-1}} \cdots \Pi_{V_1}\mathbf{x} - \Pi_V\mathbf{x}\|_2 \leq z\|\mathbf{x}\|_2, \quad \text{for all } \mathbf{x}, \tag{3.16}$$

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. $\square$

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:*

$$|\cap\pi(\mathcal{G})| \leq |\cap\mathcal{G}|, \quad \textit{for } \mathcal{G} \in \binom{\mathcal{H}}{r} \cup \binom{\mathcal{H}}{r+1}, \tag{3.17}$$

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 (3.17), 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 (3.17)) $|\cap \mathcal{G}_1| \geq 1$. 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 (3.17).

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. Let us begin by showing that $\dim(\boldsymbol{\mathcal{B}}_{\pi(S)}) = \dim(\boldsymbol{\mathcal{A}}_S)$ for all $S \in \mathcal{H}$. Note that since $\|\mathbf{A}\mathbf{x}\|_2 \leq \max_j \|\mathbf{A}_j\|_2\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_1 \leq \sqrt{k}\|\mathbf{x}\|_2$ for all k-sparse $\mathbf{x}$, by (2.1) 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 (3.7) is less than one. From (3.5), 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 shall next demonstrate that (3.17) 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 (3.5) 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 $S \in \mathcal{H}$.[6] 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 (3.4), since trivially $\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})} \subseteq \cap\boldsymbol{\mathcal{B}}_{\pi(\mathcal{G})}$ and also $\boldsymbol{\mathcal{A}}_{\cap\mathcal{G}} = \cap\boldsymbol{\mathcal{A}}_{\mathcal{G}}$ 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 (3.4) the numerator of each term in the sum above is bounded by $d\left(\boldsymbol{\mathcal{B}}_{\pi(S)}, \boldsymbol{\mathcal{A}}_S\right) = d\left(\boldsymbol{\mathcal{A}}_S, \boldsymbol{\mathcal{B}}_{\pi(S)}\right) \leq \varepsilon$,

[6] 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.

with the equality from (3.6) since $\dim(\boldsymbol{\mathcal{B}}_{\pi(S)}) = \dim(\boldsymbol{\mathcal{A}}_S)$. Thus, altogether:

$$d(\boldsymbol{\mathcal{B}}_{\cap\pi(\mathcal{G})}, \boldsymbol{\mathcal{A}}_{\cap\mathcal{G}}) \leq \frac{|\mathcal{G}|\varepsilon}{1 - \xi(\boldsymbol{\mathcal{A}}_{\mathcal{G}})} \leq \frac{C_2\varepsilon}{\max_j \|\mathbf{A}_j\|_2}, \tag{3.18}$$

recalling the definition of C_2 in (3.10). 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 (3.17) 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 (3.6) and (3.18) 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 (3.2), with J and $\overline{\varepsilon}$ in place of $[m]$ and ε, respectively. The same arguments of the case $k = 1$ in Sec. 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 $\mathbf{e}_{\overline{\pi}(i)}$ 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 $i \in J$. $\square$

Chapter 4

Discussion

The main motivation for this work was the observation that characteristic sparse representations tend to emerge from sparse coding models trained over a variety of natural scene datasets by a variety of learning algorithms. The theorems proven here provide some insight into this phenomenon by establishing very general conditions under which identification of the model parameters is not only possible but also robust to measurement and modeling error.

The guarantees concerning the identification of a dictionary and corresponding sparse codes of minimal average support size (Thm. 2), which is the optimization problem of most interest to practitioners (Prob. 2), are to my knowledge the first of their kind in both the noise-free and noisy domains. It has been shown here that, given sufficient data, this problem reduces to an instance of Prob. 1 to which the main result (Thm. 1) then applies: every dictionary and corresponding set of sparse codes consistent with the data are equivalent up to inherent relabeling/scaling ambiguities and a discrepancy (error) that scales linearly with the noise. In fact, in almost all cases these problems are well-posed given a sufficient amount of data (Thm. 3 and Cor. 2). Furthermore, the derived scaling constants are explicit and computable; as such, there is an effective procedure that suffices to affirm if a proposed solution to these problems is indeed unique up to noise and inherent ambiguities, although it is not efficient in general.

While the extension from exact recovery to the noisy stability of dictionary learning may be 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. Consequently, the inferences of those applying dictionary learning methods to inverse problems in their research are justified only in principle; but this is unavoidably the case for NP-hard problems. What sets the main results of this work apart from the vast majority of results in the field, however, is their deterministic nature. 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 (e.g., Cor. 2 makes a sweeping statement applicable to all continuous distributions).

Indeed, theoretical validation makes little practical difference if the methodology is already in widespread use, while practical criteria establishing whether the data models obtained by practitioners are optimal or not would have very high impact. To this end, the work has been laid out for those wanting to derive statistical criteria for inference with respect to more domain-specific parametric dictionaries and codes (i.e. estimate C_1), and reduced by half for those hoping to prove the consistency of any dictionary learning algorithm (i.e. prove convergence to within $\varepsilon(\delta_1, \delta_2)$ given in (2.4)).

Nonetheless, 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 these findings. Several groups have applied compressed sensing to signal processing tasks; for instance, in MRI analysis [36], image compression [13], and even the design of an ultrafast camera [17]. It is only a matter of time before these systems incorporate dictionary learning to encode and decode signals (e.g., in a device that learns structure from motion [32]), just as scientists have used sparse coding to uncover latent structure in data (e.g., forgery detection [25, 39]; brain recordings [30, 1, 33]; and gene expression [49]). As uniqueness guarantees with minimal assumptions apply to all areas of data science and engineering that utilize learned sparse structure, assurances offered by these theorems give hope that different devices and algorithms may learn equivalent representations given enough data from statistically identical systems.[1]

Within the field of theoretical neuroscience in particular, dictionary learning for sparse coding and related methods have recovered characteristic components of natural images [38, 27, 7, 22] and sounds [6, 44, 9] that reproduce response properties of cortical neurons. The results of this work suggest that this correspondence could be due to the "universality" of sparse representations in natural data, an early mathematical idea in neural theory [41]. Furthermore, they justify the soundness of one of the few hypothesized theories of bottleneck communication in the brain [29]: 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., [42, 43, 40]).[2]

4.1 Future directions

There are many challenges left open by this work. First and foremost, it should be stressed that all conditions stated here which guarantee the uniqueness and stability of sparse representations have only been shown sufficient; it remains open to work out a set of necessary conditions on all fronts, be it on the number of required samples per support, the structure of support set hypergraphs, or the tolerable signal-to-noise ratio for a bounded recovery error. It is also worth stressing that the deterministic conditions derived here must

[1]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 [19].

[2]We refer the reader to [16] for more on interactions between dictionary learning and neuroscience.

accommodate always the worst possible cases. It would be of great practical benefit to see how drastically all conditions can be relaxed by requiring less-than-certain guarantees, as (for instance) exhibited in the discussion on probabilistic pigeonholing following the proof of Thm. 1. In a similar vein, the tolerable signal-to-noise ratio can be reduced by considering the probability that noise sampled from a concentrated isotropic distribution will point in a harmful direction, which may be especially low in high-dimensional spaces or for certain support set hypergraphs.

Another interesting remaining challenge is to work out for which special cases it is efficient to check that a solution to Prob. 1 or 2 is unique up to noise and inherent ambiguities. Considering that the sufficient conditions detailed here are in general NP-hard to compute, are the necessary conditions also hard to compute? Are Probs. 1 and 2 then also hard (e.g., see [46])? Since Prob. 2 is intractable in general (i.e. including the noiseless case), but solvable in practice by convex relaxation when the matrix $\mathbf{A}$ is known and has a large enough lower bound over sparse domains [14], is there a version of Thm. 2 that lays down general conditions under which Prob. 2 can be solved efficiently in full by similar means?

I briefly expand on some of these directions below. It is my hope that these remaining challenges pique the interest of the community and that the theoretical tools showcased here can be applied to derive practical dictionary learning guidelines for sparse representation modeling of real data.

Reducing the required signal-to-noise ratio

A concern raised in peer review of this work was the typical size of the constant C_1, which sets the tolerable signal-to-noise ratio for dictionary and code recovery up to an acceptable error. Referring to the definition of this constant in (3.11), the reader should note that the denominator involves L_k, a standard quantity in the field of compressed sensing (the "restricted isometry constant", see footnote 3), which is known to be reasonable for many random distributions generating dictionaries $\mathbf{A}$ and sparse codes $\mathbf{x}_i$ [4]. The numerator C_2, on the other hand, incorporates the more obscure quantity ξ defined in (4), which is computed from the "Friedrichs angle" between certain spans of subsets of the columns of $\mathbf{A}$. Simulations for small (pseudo-)randomly generated dictionaries $\mathbf{A}$ suggest nonetheless that the constant C_2 is likely reasonable in general as well (at least, for the case where $m = k^2$ and $\mathcal{H}$ is taken to be the set of rows and columns formed by arranging the elements of $[m]$ into a square grid; see Fig. 4.1). These observations motivate the following conjecture:

Conjecture 1. *For all $t > 0$,*

$$Pr[|C_2 - \mathbb{E}[C_2]| > t] \to 0 \quad \textit{as } k \to \infty \textit{ and } k/m \to 0$$

provided the assumptions of Thm. 1 are satisfied.

Reducing the required number of samples

It is possible to tighten the pigeonholing argument in the proof of Thm. 1 and thereby reduce the deterministic sample complexity without recourse to uncertain guarantees. The argument as presented iterates over supports $S \in \mathcal{H}$, in each case determining a corresponding support $\overline{S} \in \binom{[\overline{m}]}{k}$ without consideration of previously matched support pairs; and yet the assumption $L_{2\mathcal{H}}(\mathbf{A}) > 0$ implies that no two supports in $\mathcal{H}$ can map to the same $\overline{S}$. The number of bins to pigeonhole into thus decreases every iteration, though this is a drop in a bucket of exponential size. It would be interesting to see how much the deterministic sample complexity can be reduced by imposing these constraints holistically, given the specific structure of the hypergraph $\mathcal{H}$.

Incidentally, there is also room to breathe in the restrictions on $\mathcal{H}$. Already, the results of this work motivate the following question, which is only one among many combinatorial problems brought to mind by the SIP (Def. 2):

Question 1. *Fix integers m and $k < m$. What is the smallest regular hypergraph $\mathcal{H} \subseteq \binom{[m]}{k}$ satisfying the SIP?*

A close examination of the proof of Lemma 4 and its incorporation into the proof of Lem. 1 reveals, however, that looser constraints on $\mathcal{H}$ maintain compatibility with the argument to some degree. Assuming that the nodes of $\mathcal{H}$ are labeled in order of non-increasing degree, if instead of the full SIP we require only that the stars $\cap\sigma(i)$ intersect at singletons for all $i \leq q$ for some positive q, we have that $\overline{m} \geq k|\mathcal{H}|/\deg(1)$ and, provided $\overline{m} < k|\mathcal{H}|/(\deg(1) - 1)$, the nonempty subset J is of size equal to the largest number p satisfying:

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

Specifically, J contains all nodes of degree exceeding $\deg(p)$ and some subset of those with degree equal to $\deg(p)$. It is thus natural to ask: what are the necessary constraints on the hypergraph $\mathcal{H}$, and what is the smallest hypergraph satisfying these constraints for given m and k?

Opening ourselves up instead to uncertain guarantees, we can ask:

Question 2. *Fix $k < m$. What is the probability that a random subset of $\binom{[m]}{k}$ is regular and satisfies the SIP?*

We may also elaborate on the probabilistic pigeonholing strategy outlined in Sec. 3 following the proof of Thm. 1. The problem is to count the number of ways in which vectors supported in $S \in \mathcal{H}$ can be partitioned among supports in $\binom{[\overline{m}]}{k}$ without allocating k or more to any individual one (in which case the logic of the proof fails to imply the result; we are interested in the probability that it doesn't). These are integer solutions to the problem $\sum_i n_i = N$ subject to $n_i < k$ for all i, where $i = 1, \ldots, \binom{\overline{m}}{k}$. Following closely the exposition in [34], it appears there is no closed formula for this problem, but the number of solutions can be computed in a number of operations independent of N. Writing $p = \binom{m}{k}$, the number is the coefficient of X^N in the polynomial $(1 + X + \ldots + X^{k-1})^p$. Written as a rational function of X,

$$(1 + X + \ldots + X^{k-1})^p = \left(\frac{1 - X^k}{1 - X}\right)^p = \frac{\left(1 - X^k\right)^p}{(1 - X)^p}$$

the coeffiecient of X^i in the numerator is zero unless i is a multiple qk of k, in which case it is $(-1)^q\binom{p}{q}$, and the coefficient of X^j in the inverse of the denominator is $(-1)^j\binom{-p}{j} = \binom{j+p-1}{j}$, which is zero unless $j \geq 0$ and otherwise equal to $\binom{j+p-1}{p-1}$. It remains to sum over all $i + j = N$, which gives:

$$n_{\text{fails}} = \sum_{q=0}^{\min(p,N/k)} (-1)^q \binom{p}{q} \binom{N - qk + p - 1}{p - 1}$$

where the summation is truncated to ensure that $N - qk \geq 0$ (the condition $j \geq 0$) and has at most $p + 1 = \binom{m}{k} + 1$ terms.

The total number of ways to pigeonhole is $n_{\text{total}} = \binom{N+p-1}{p-1}$, and so the probability of full recovery is $(1 - n_{\text{fails}}/n_{\text{total}})^{|\mathcal{H}|}$. Curves computed in this way in (see Fig. 4.1) suggest that while it may very well be impossible to exorcise exponentiality from the number of required samples in the deterministic or almost-certain case, perhaps it is possible with high-probability by one way or another. Informally,

Question 3. *While fixing k yields polynomial deterministic sample complexity in m (see Cor. 1), is there some more general probablistic sense (perhaps for some restricted class of hypergraphs) by which sample complexity is polynomial in both m and k?*

Dictionary learning via ℓ_1-norm minimization

A commonly applied workaround to the intractability (see [46]) of Prob. 2 is to replace the ℓ_0-norm in (1.4) with its best convex approximation, the ℓ_1-norm, thereby transforming the inference of sparsest $\overline{\mathbf{x}}_i$ for fixed $\mathbf{B}$ into a convex optimization problem solvable by a linear program. Recalling the linear model (1.1) of data $\mathbf{z}_i$, the approach is to solve the following problem rather than solving Prob 2 directly.:

Problem 3. *Find a matrix $\mathbf{B}$ and vectors $\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_N$ solving:*

$$\min \sum_{i=1}^{N} \|\overline{\mathbf{x}}_i\|_1 \quad \textit{subject to} \quad \|\mathbf{z}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \eta, \ \textit{for all } i, \tag{4.1}$$

A major advance in compressive sensing was the discovery of practical criteria guaranteeing that, when the matrix $\mathbf{B}$ is held fixed, the vectors $\overline{\mathbf{x}}_i$ that solve (4.1) also in fact solve (1.4) provided $L_{2k}(\mathbf{B})$ is large enough [14]. Interestingly, the current work provides conditions on the generating dictionary $\mathbf{A}$ and k-sparse codes $\mathbf{x}_i$ under which *all* matrices $\mathbf{B}$ solving Prob. 2 have $L_{2k}(\mathbf{B})$ bounded from below; specifically, $L_{2k}(\mathbf{B}) \geq (L_{2k}(\mathbf{A}) - 2\eta C_1) / \|\mathbf{D}\|_1$ in the case where $\mathbf{A}$ satisfies the spark condition (1.2). Thus, for suitable $\mathbf{A}$ there should be some noise bound inside of which all solutions to Prob. 2 are solutions to Prob. 3 as well. It is an open question as to whether there exist practical constraints which reject solutions to Prob. 3 that don't also solve Prob. 2.

We can already see this may be possible by examining the simple case $k = 1$ without noise, i.e. $\eta = 0$.

Proposition 1. *If the matrix* $\mathbf{A}$ *has unit length-columns with* $L_2(\mathbf{A}) > 0$ *and* $\mathbf{x}_i = c_i \mathbf{e}_i$ *with* $c_i \neq 0$ *for* $i = 1, \ldots, m$*, then every solution to Prob. 3 with* $\mathbf{B}$ *also having* m *unit-length columns satisfies* $\mathbf{A} = \mathbf{BP}$ *and* $\mathbf{x}_i = \mathbf{P}^\top \overline{\mathbf{x}}_i$ *for some* $m \times m$ *permutation matrix* $\mathbf{P}$.

Proof. We will show that the $\overline{\mathbf{x}}_i$ must in fact be 1-sparse and then apply Thm. 1. Fixing $i \in [m]$ and writing $\overline{\mathbf{x}}_i = \sum_{j=1}^m \overline{c}_j \mathbf{e}_j$, we have:

$$c_i = \|\mathbf{A}\mathbf{x}_i\|_2 = \|\mathbf{B}\overline{\mathbf{x}}_i\|_2 = \|\sum_{j=1}^m \overline{c}_j \mathbf{B}\mathbf{e}_j\|_2 \leq \sum_{j=1}^m |\overline{c}_j| \|\mathbf{B}\mathbf{e}_j\|_2 = \|\overline{\mathbf{x}}_i\|_1 \tag{4.2}$$

So $\|\overline{\mathbf{x}}_i\|_1 \geq c_i$ for all $i \in [m]$, and we have $\sum_{i=1}^m \|\overline{\mathbf{x}}_i\|_1 \geq \sum_{i=1}^m c_i$. But since $\mathbf{A}$ and the $\mathbf{x}_i$ satisfy the constraints of the minimization problem, we must also have $\sum_{i=1}^m \|\overline{\mathbf{x}}_i\|_1 \leq \sum_{i=1}^m \|\mathbf{x}_i\|_1 = \sum_{i=1}^m c_i$. Thus, $\sum_{i=1}^m \|\overline{\mathbf{x}}_i\|_1 = \sum_{i=1}^m c_i$. Since again $\|\overline{\mathbf{x}}_i\|_1 \geq c_i$, it must be the case that in fact $\|\overline{\mathbf{x}}_i\|_1 = c_i$ for all $i \in [m]$.

Revisiting (4.2), for a given $i \in [m]$ we have $c_i = \|\mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \|\overline{\mathbf{x}}_i\|_1 = c_i$, with equality holding only when the vectors $\overline{c}_j \mathbf{B}\mathbf{e}_j$ are colinear. This would be the case either if $\overline{\mathbf{x}}_i$ is 1-sparse, or if $\mathbf{B}$ has colinear columns. We can rule out the latter case because if that were so, we could consolidate all sets of colinear columns to form a dictionary with fewer than m columns and yet for which there must exist 1-sparse $\overline{\mathbf{x}}_i$, $i = 1, \ldots, m$ solving Prob. 3; but this would contradict the fact that $L_2(\mathbf{A}) > 0$. So it must be the case that in fact the $\overline{\mathbf{x}}_i$ are all 1-sparse, and we may apply Thm. 1. □

The approach taken here was to show that the points $\mathbf{B}\overline{\mathbf{x}}_i$ lie on a polytype which identifies with all of the data $\mathbf{z}_i = c_i \mathbf{A}\mathbf{e}_i$ only at its vertices, and that the $\overline{\mathbf{x}}_i$ must therefore be 1-sparse. Intuitively, it seems that similar arguments should apply in the general case $k < m$, where the points $\mathbf{B}\overline{\mathbf{x}}_i$ may align with all of the data only at the $k-1$-dimensional boundaries of the polytope.

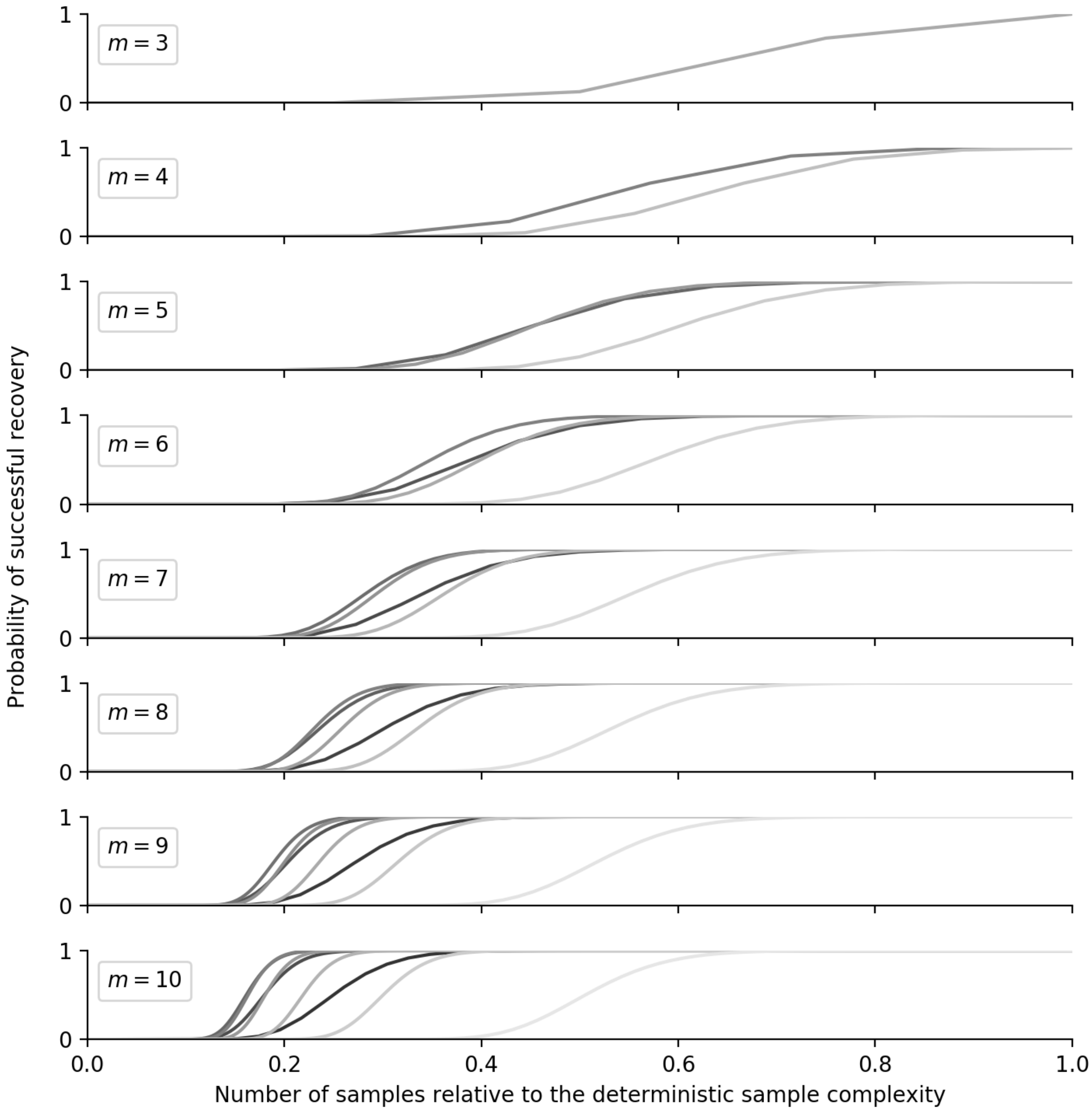

Figure 4.1: **Learning a dictionary from an arbitrary number of samples.** Probability of successful dictionary and code recovery (as per Thm. 1) for a number of samples N given as a fraction of the deterministic sample complexity $N = |\mathcal{H}|[(k-1)\binom{m}{k} + 1]$ when the support set hypergraph $\mathcal{H}$ is the set of m consecutive intervals of length k in a cyclic order on $[m]$. Each plot has k ranging from 2 to $m-1$ (the case $k = 1$ requires $N = m$), with lighter grey lines corresponding to larger k. Successful recovery is nearly certain with far fewer samples than the deterministic sample complexity.

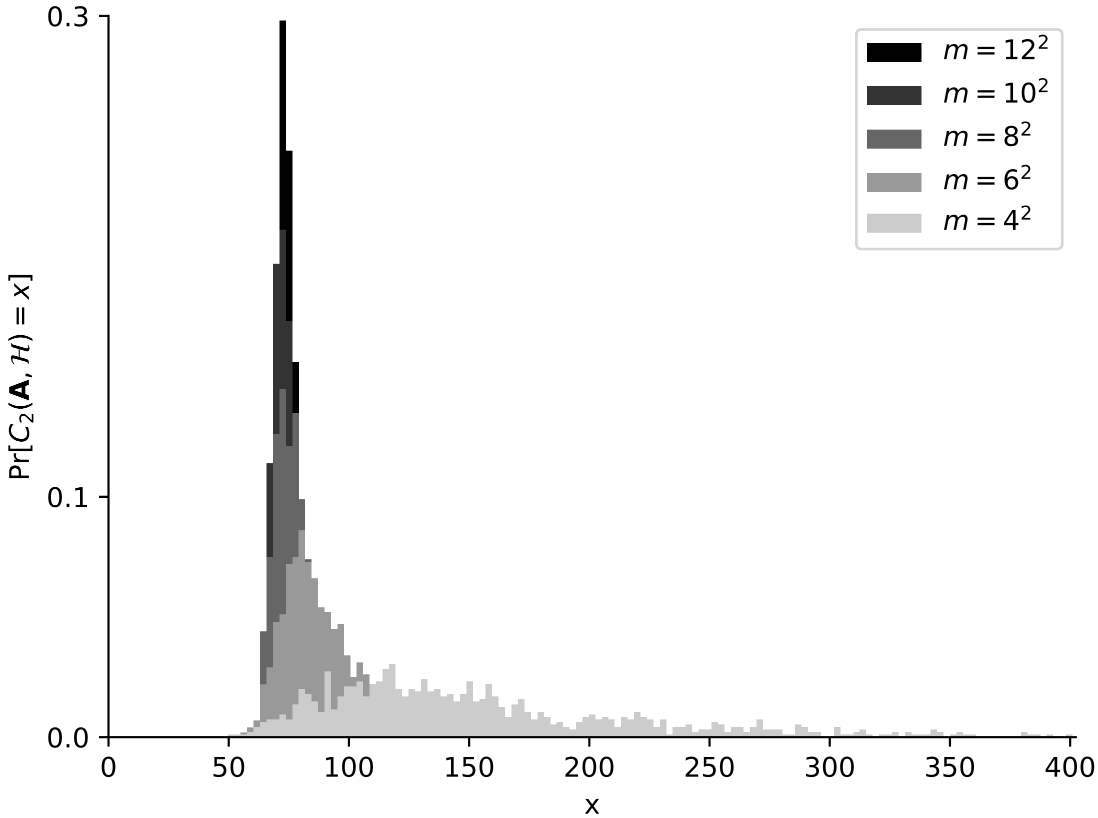

Figure 4.2: **Concentration of the constant C_2.** Distribution of $C_2(\mathbf{A}, \mathcal{H})$ computed for 1.33x overcomplete generic unit-norm dictionaries $\mathbf{A} \in \mathbb{R}^{n\times m}$ (i.e. with $n = 3m/4$) when the support set hypergraph $\mathcal{H}$ consists of the rows and columns formed by arranging the elements of $[m]$ into a square grid (i.e. $m = k^2$). The distribution becomes more concentrated as m grows.

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] Richard Baraniuk et al. “A simple proof of the restricted isometry property for random matrices”. In: *Constructive Approximation* 28.3 (2008), pp. 253–263.

[5] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in real algebraic geometry.* Vol. 10. Springer, 2006.

[6] A. Bell and T. Sejnowski. “Learning the higher-order structure of a natural sound”. In: *Network Comp. Neural* 7.2 (1996), pp. 261–266.

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

[8] 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.

[9] N. Carlson, V. Ming, and M. DeWeese. “Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus”. In: *PLOS Comput. Biol.* 8.7 (2012), e1002594. DOI: `10.1371/journal.pcbi.1002594`.

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

[11] 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.

[12] 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.

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

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

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

[16] 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.

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

[18] 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.

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

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

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

[22] 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.

[23] C. Hillar. (personal communication).

[24] 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.

[25] 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.

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

[27] A. Hyvarinen. “Fast and robust fixed-point algorithms for independent component analysis”. In: *IEEE transactions on Neural Networks* 10.3 (1999), pp. 626–634.

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

[29] 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.

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

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

[32] 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.

[33] 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.

[34] Marc van Leeuwen. *Answer to: extended stars-and-bars problem(where the upper limit of the variable is bounded).* 2013 (accessed July 18, 2020). URL: `https://math.stackexchange.com/questions/553960/extended-stars-and-bars-problemwhere-the-upper-limit-of-the-variable-is-bounded`.

[35] 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.

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

[37] 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.

[38] 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.

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

[40] 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.

[41] W. Pitts and W. McCulloch. "How we know universals: the perception of auditory and visual forms". In: *Bull. Math. Biophys.* 9.3 (1947), pp. 127–147.

[42] 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.

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

[44] E.C. Smith and M.S. Lewicki. "Efficient auditory coding". In: *Nature* 439.7079 (2006), pp. 978–982.

[45] 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.

[46] A. Tillmann. "On the computational intractability of exact and approximate dictionary learning". In: *IEEE Signal Process. Lett.* 22.1 (2015), pp. 45–49.

[47] 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.

[48] 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.

[49] 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.

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