On the Uniqueness and Stability of Dictionaries for Sparse Representation of Noisy Signals

Charles J. Garfinkle and Christopher J. Hillar

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

***Index Terms*—Inverse problems, brain modeling, parameter estimation, sparse matrices, unsupervised learning, channel models.**

I. Introduction

SPARSE coding is a common modern approach to pattern analysis in signal processing whereby each of N observed n-dimensional signal samples is viewed as a (noisy) linear combination of at most k elementary waveforms drawn from some unknown "dictionary" of size $m \ll N$ (see [1] for a comprehensive review). Optimizing dictionaries subject to this and related sparsity constraints has revealed seemingly characteristic sparse structure in several signal classes of current interest (e.g., in vision [2]).

Of particular note are the seminal works in the field [3]–[6], which demonstrated that dictionaries optimized for coding small patches of "natural" images share qualitative similarities with linear filters estimated from response properties of simple-cell neurons in mammalian visual cortex. Curiously, these waveforms (e.g., "Gabor" wavelets) appear in dictionaries learned by a variety of algorithms trained over different natural image datasets, suggesting that these learned features may, in some sense, be canonical [7].

Manuscript received December 24, 2018; revised May 10, 2019; accepted June 18, 2019. Date of publication August 19, 2019; date of current version November 7, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dr. Athanasios A. Rontogiannis. This work was supported in part by NSF under Grant IIS-1219212. *(Corresponding author: Christopher J. Hillar.)*

The authors are with the Redwood Center for Theoretical Neuroscience, Berkeley, CA 94720, USA (e-mail: cjfinks@gmail.com; chillar@msri.org).

Digital Object Identifier 10.1109/TSP.2019.2935914

Motivated by these discoveries and more recent work relating compressed sensing [8] to a theory of information transmission through random wiring bottlenecks in the brain [9], we address when dictionaries for sparse representation are indeed identifiable from data. Answers to this question may also have implications in practice wherever an appeal is made to latent sparse structure of data (e.g., forgery detection [10], [11]; brain recordings [12]–[14]; and gene expression [15]).

While several algorithms have been recently proposed to provably recover unique dictionaries under specific conditions (see [16, Section I-E] for a summary of the state-of-the-art), few theorems can be invoked to justify the consistency of inference under this model of data more broadly. To our knowledge, a universal guarantee of the uniqueness and stability of learned dictionaries and the sparse representations they induce over noisy data has yet to appear in the literature.

Here, we prove very generally that uniqueness and stability is a typical property of sparse dictionary learning. More specifically, we show that matrices injective on a sparse domain are identifiable from $N = m(k-1)\binom{m}{k} + m$ noisy linear combinations of k of their m columns up to an error that is linear in the noise (Theorem 1 and Corollary 1). In fact, provided $n \geq \min(2k, m)$, in almost all cases the problem is well-posed, as per Hadamard [17], given a sufficient amount of data (Theorem 3 and Corollary 2).

Our guarantees also extend to a related (and perhaps more commonly posed, e.g. [18]) optimization problem seeking a dictionary minimizing the average number of elementary waveforms required to reconstruct each sample of the dataset (Theorem 2). To practical benefit, our results impose no restrictions on learned dictionaries (e.g., that they, too, be injective over some sparse domain) beyond an upper bound on dictionary size, which is necessary in any case to avoid trivial solutions (e.g., allowing $m = N$).

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

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

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

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