The Sample Complexity of Dictionary Learning

Daniel Vainsencher DANIELV@TX.TECHNION.AC.IL
Shie Mannor SHIE@EE.TECHNION.AC.IL
Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa 32000, Israel

Alfred M. Bruckstein FREDDY@CS.TECHNION.AC.IL
Department of Computer Science
Technion, Israel Institute of Technology
Haifa 32000, Israel

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

A large set of signals can sometimes be described sparsely using a dictionary, that is, every element can be represented as a linear combination of few elements from the dictionary. Algorithms for various signal processing applications, including classification, denoising and signal separation, learn a dictionary from a given set of signals to be represented. Can we expect that the error in representing by such a dictionary a previously unseen signal from the same source will be of similar magnitude as those for the given examples? We assume signals are generated from a fixed distribution, and study these questions from a statistical learning theory perspective.

We develop generalization bounds on the quality of the learned dictionary for two types of constraints on the coefficient selection, as measured by the expected L_2 error in representation when the dictionary is used. For the case of l_1 regularized coefficient selection we provide a generalization bound of the order of $O\left(\sqrt{np\ln(m\lambda)/m}\right)$, where n is the dimension, p is the number of elements in the dictionary, λ is a bound on the l_1 norm of the coefficient vector and m is the number of samples, which complements existing results. For the case of representing a new signal as a combination of at most k dictionary elements, we provide a bound of the order $O(\sqrt{np\ln(mk)/m})$ under an assumption on the closeness to orthogonality of the dictionary (low Babel function). We further show that this assumption holds for *most* dictionaries in high dimensions in a strong probabilistic sense. Our results also include bounds that converge as $1/m$, not previously known for this problem. We provide similar results in a general setting using kernels with weak smoothness requirements.

Keywords: statistical machine learning, dictionary learning, generalization bounds, signal processing, kernel methods

1. Introduction

A common technique in processing signals from $\mathcal{X} = \mathbb{R}^n$ is to use sparse representations; that is, to approximate each signal x by a "small" linear combination a of elements d_i from a dictionary $D \in \mathcal{X}^p$, so that $x \approx Da = \sum_{i=1}^{p} a_i d_i$. This has various uses detailed in Section 1.1. The smallness of a is often measured using either $\|a\|_1$, or the number of non zero elements in a, often denoted $\|a\|_0$. The approximation error is measured here using a Euclidean norm appropriate to the vector space. We denote the approximation error of x using dictionary D and coefficients from a set A by

$$h_{A,D}(x) = \min_{a \in A} \|Da - x\|\,, \tag{1.1}$$

where A is one of the following sets determining the sparsity required of the representation:

$$H_k = \{a : \|a\|_0 \leq k\}$$

induces a "hard" sparsity constraint, which we also call k sparse representation, while

$$R_\lambda = \{a : \|a\|_1 \leq \lambda\}$$

induces a convex constraint that is considered a "relaxation" of the previous constraint.

The dictionary learning problem is to find a dictionary D minimizing

$$E(D) = \mathbb{E}_{x \sim \nu} h_{A,D}(x), \tag{1.2}$$

where ν is a distribution over signals that is known to us only through samples from it. The problem addressed in this paper is the "generalization" (in the statistical learning sense) of dictionary learning: to what extent does the performance of a dictionary chosen based on a finite set of samples indicate its expected error in (1.2)? This clearly depends on the number of samples and other parameters of the problem such as the dictionary size. In particular, an obvious algorithm is to represent each sample using itself, if the dictionary is allowed to be as large as the sample, but the performance on unseen signals is likely to disappoint.

To state our goal more quantitatively, assume that an algorithm finds a dictionary D suited to k sparse representation, in the sense that the average representation error $E_m(D)$ on the m examples given to the algorithm is low. Our goal is to bound the generalization error ε, which is the additional expected error that might be incurred:

$$E(D) \leq (1 + \eta) E_m(D) + \varepsilon, \tag{1.3}$$

where $\eta \geq 0$ is sometimes zero, and the bound ε depends on the number of samples and problem parameters. Since efficient algorithms that find the optimal dictionary for a given set of samples (also known as empirical risk minimization, or ERM, algorithms) are not known for dictionary learning, we prove uniform convergence bounds that apply simultaneously over all admissible dictionaries D, thus bounding from above the sample complexity of the dictionary learning problem. In particular, such a result means that every procedure for approximate minimization of empirical error (empirical dictionary learning) is also a procedure for approximate dictionary learning (as defined above) in a similar sense.

Many analytic and algorithmic methods relying on the properties of finite dimensional Euclidean geometry can be applied in more general settings by applying kernel methods. These consist of treating objects that are not naturally represented in $\mathbb{R}^n$ as having their similarity described by an inner product in an abstract *feature space* that is Euclidean.

This allows the application of algorithms depending on the data only through a computation of inner products to such diverse objects as graphs, DNA sequences and text documents (Shawe-Taylor and Cristianini, 2004). Is it possible to extend the usefulness of dictionary learning techniques to this setting? We address sample complexity aspects of this question as well.

1.1. Background and related work

Sparse representations are by now standard practice in diverse fields such as signal processing, natural language processing, etc. Typically, the dictionary is assumed to be known. The motivation for sparse representations is indicated by the following results, in which we assume the signals come from $\mathcal{X} = \mathbb{R}^n$ are normalized to have length 1, and the representation coefficients are constrained to $A = H_k$ where $k < n, p$ and typically $h_{A,D}(x) \ll 1$.

- Compression: If a signal x has an approximate sparse representation in some commonly known dictionary D, it can be stored or transmitted more economically with reasonable precision. Finding a good sparse representation can be computationally hard but if D fulfills certain geometric conditions, then its sparse representation is unique and can be found efficiently (see, e.g., Bruckstein et al., 2009).
- Denoising: If a signal x has a sparse representation in some known dictionary D, and $\tilde{x} = x + \nu$, where the random noise ν is Gaussian, then the sparse representation found for $\tilde{x}$ will likely be very close to x (for example Chen et al., 2001).
- Compressed sensing: Assuming that a signal x has a sparse representation in some known dictionary D that fulfills certain geometric conditions, this representation can be approximately retrieved with high probability from a small number of random linear measurements of x. The number of measurements needed depends on the sparsity of x in D (Candes and Tao, 2006).

The implications of these results are significant when a dictionary D is known that sparsely represents simultaneously many signals. In some applications the dictionary is chosen based on prior knowledge, but in many applications the dictionary is learned based on a finite set of examples. To motivate dictionary learning, consider an image representation used for compression or denoising. Different types of images may have different properties (MRI images are not similar to scenery images), so that learning a dictionary specific to each type of images may lead to improved performance. The benefits of dictionary learning have been demonstrated in many applications (Protter and Elad, 2007; Peyré, 2009).

Two extensively used techniques related to dictionary learning are Principal Component Analysis (PCA) and K-means clustering. The former finds a single subspace minimizing the sum of squared representation errors which is very similar to dictionary learning with $A = H_k$ and $p = k$. The latter finds a set of locations minimizing the sum of squared distances between each signal and the location closest to it which is very similar to dictionary learning with $A = H_1$ where p is the number of locations. Thus we could see dictionary learning as PCA with multiple subspaces, or as clustering where multiple locations are used to represent each signal. The sample complexities of both algorithms are well studied (Bartlett et al., 1998; Biau et al., 2008; Shawe-Taylor et al., 2005; Blanchard et al., 2007).

This paper does not address questions of computational cost, though they are very relevant. Finding optimal coefficients for k sparse representation (that is, minimizing (1.1) with $A = H_k$) is NP-hard in general (Davis et al., 1997). Dictionary learning as the optimization problem of minimizing (1.2) is less well understood, even for empirical ν (consisting of a finite number of samples), despite over a decade of work on related algorithms with good empirical results (Olshausen and Fieldt, 1997; Lewicki et al., 1998; Kreutz-Delgado et al., 2003; Aharon et al., 2006; Lee et al., 2007; Mairal et al., 2010).

The only prior work we are aware of that addresses generalization in dictionary learning, by Maurer and Pontil (2010), addresses the convex representation constraint $A = R_\lambda$; we discuss the relation of our work to theirs in Section 2.

2. Results

Except where we state otherwise, we assume signals are generated in the unit sphere $\mathbb{S}^{n-1}$. Our results are:

A new approach to dictionary learning generalization. Our first main contribution is an approach to generalization bounds in dictionary learning that is complementary to the approach used by Maurer and Pontil (2010). The previous result, given below in Theorem 6 has generalization error bounds of order

$$O\left(\sqrt{p\min(p,n)\left(\lambda+\sqrt{\ln(m\lambda)}\right)^2/m}\right)$$

on the squared representation error. A notable feature of this result is the weak dependence on the signal dimension n. In Theorem 1 we quantify the complexity of the class of functions $h_{A,D}$ over all dictionaries whose columns have unit length, where $A \subset R_\lambda$. Combined with standard methods of uniform convergence this results in generalization error bounds ε of order $O\left(\sqrt{np\ln(m\lambda)/m}\right)$ when $\eta = 0$. While our bound does depend strongly on n, this is acceptable in the case $n < p$, also known in the literature as the "over-complete" case (Olshausen and Fieldt, 1997; Lewicki et al., 1998). Note that our generalization bound applies with different constants to the representation error itself and many variants including the squared representation error, and has a weak dependence on λ. The dependence on λ is significant, for example, when $\|a\|_1$ is used as a weighted penalty term by solving $\min_a \|Da - X\| + \gamma \cdot \|a\|_1$; in this case $\lambda = O\left(\gamma^{-1}\right)$ may be quite large.

Fast rates. For the case $\eta > 0$ our methods allow bounds of order $O(np\ln(\lambda m)/m)$. The main significance of this is in that the general statistical behavior they imply occurs in dictionary learning. For example, generalization error has a "proportional" component which is reduced when the empirical error is low. Whether fast rates results can be proved in the dimension free regime is an interesting question we leave open. Note that due to lower bounds by Bartlett et al. (1998) of order $\sqrt{m^{-1}}$ on the k-means clustering problem, which corresponds to dictionary learning for 1-sparse representation, fast rates may be expected only with $\eta > 0$, as presented here.

We now describe the relevant function class and the bounds on its complexity, which are proved in Section 3. The resulting generalization bounds are given explicitly at the end of this section.

Theorem 1 *For every $\varepsilon > 0$, the function class*

$$\mathcal{G}_\lambda = \left\{ h_{R_\lambda,D} : \mathbb{S}^{n-1} \to \mathbb{R} : D \in \mathbb{R}^{n\times p}, \|d_i\| \leq 1 \right\},$$

taken as a metric space with the distance induced by $\|\cdot\|_\infty$, has a subset of cardinality at most $(4\lambda/\varepsilon)^{np}$, such that every element from the class is at distance at most ε from the subset.

While we give formal definitions in Section 3, such a subset is called an ε cover, and such a bound on its cardinality is called a covering number bounds.

Extension to k sparse representation. Our second main contribution is to extend both our approach and that of Maurer and Pontil (2010) to provide generalization bounds for dictionaries for k sparse representations, by using a bound λ on the l_1 norm of the representation coefficients when the dictionaries are close to orthogonal. Distance from orthogonality is measured by the Babel function (which, for example, upper bounds the magnitude of the maximal inner product between distinct dictionary elements) defined below and discussed in more detail in Section 4.

Definition 2 (Babel function, Tropp 2004) *For any $k \in \mathbb{N}$, the Babel function $\mu_k : \mathbb{R}^{n\times m} \to \mathbb{R}^+$ is defined by:*

$$\mu_k(D) = \max_{i\in\{1,\ldots,p\}} \max_{\Lambda\subset\{1,\ldots,p\}\setminus\{i\};|\Lambda|=k} \sum_{j\in\Lambda} |\langle d_j, d_i\rangle| .$$

The following proposition, which is proved in Section 3, bounds the 1-norm of the dictionary coefficients for a k sparse representation and also follows from analysis previously done by Donoho and Elad (2003); Tropp (2004).

Proposition 3 *Let each column d_i of D fulfill $\|d_i\| \in [1,\gamma]$ and $\mu_{k-1}(D) \leq \delta < 1$, then a coefficient vector $a \in \mathbb{R}^p$ minimizing the k-sparse representation error $h_{H_k,D}(x)$ exists which has $\|a\|_1 \leq \gamma k/(1-\delta)$.*

We now consider the class of all k sparse representation error functions. We prove in Section 3 the following bound on the complexity of this class.

Corollary 4 *The function class*

$$\mathcal{F}_{\delta,k} = \left\{ h_{H_k,D} : \mathbb{S}^{n-1} \to \mathbb{R} : \mu_{k-1}(D) < \delta, d_i \in \mathbb{S}^{n-1} \right\},$$

taken as a metric space with the metric induced by $\|\cdot\|_\infty$, has a covering number bound of $(4k/(\varepsilon(1-\delta)))^{np}$.

The dependence of the last two results on $\mu_{k-1}(D)$ means that the resulting bounds will be meaningful only for algorithms which explicitly or implicitly prefer near orthogonal dictionaries. Contrast this to Theorem 1 which has no significant conditions on the dictionary.

Asymptotically almost all dictionaries are near orthogonal. A question that arises is what values of μ_{k-1} can be expected for parameters n, p, k? We shed some light on this question through the following probabilistic result, which we discuss in Section 4 and prove in the full version.

Theorem 5 *Suppose that D consist of p vectors chosen uniformly and independently from $\mathbb{S}^{n-1}$. Then we have*

$$P\left(\mu_k > \frac{1}{2}\right) \leq \frac{1}{\left(e^{(n-2)/(10k\ln p)^2} - 1\right)}.$$

Since low values of the Babel function have implications to representation finding algorithms, this result is of interest also outside the context of dictionary learning. Essentially it means that random dictionaries of size sub-exponential in $(n-2)/k^2$ have low Babel function.

New generalization bounds for l_1 case. The covering number bound of Theorem 1 implies several generalization bounds for the problem of dictionary learning for l_1 regularized representation which we give here. These differ from those by Maurer and Pontil (2010) in depending more strongly on the dimension of the space, but less strongly on the particular regularization term. We first give the relevant specialization of the result by Maurer and Pontil (2010) for comparison and for reference as we will later build on it. This result is independent of the dimension n of the underlying space, thus the Euclidean unit ball B may be that of a general Hilbert space, and the errors measured by $h_{A,D}$ are in the same norm.

Theorem 6 (Maurer and Pontil 2010) *Let $A \subset R_\lambda$, and let ν be any distribution on the unit sphere B. Then with probability at least $1 - e^{-x}$ over the m samples in E_m drawn according to ν, for all dictionaries $D \subset B$ with cardinality p:*

$$Eh^2_{A,D} \leq E_m h^2_{A,D} + \sqrt{\frac{p^2\left(14\lambda + 1/2\sqrt{\ln\left(16m\lambda^2\right)}\right)^2}{m}} + \sqrt{\frac{x}{2m}}.$$

Using the covering number bound of Theorem 1 and a bounded differences concentration inequality (see Lemma 21), we obtain the following result. The details are given in Section 3.

Theorem 7 *Let $\lambda > e/4$, with ν a distribution on $\mathbb{S}^{n-1}$. Then with probability at least $1 - e^{-x}$ over the m samples in E_m drawn according to ν, for all D with unit length columns:*

$$Eh_{R_\lambda,D} \leq E_m h_{R_\lambda,D} + \sqrt{\frac{np\ln\left(4\sqrt{m}\lambda\right)}{2m}} + \sqrt{\frac{x}{2m}} + \sqrt{\frac{4}{m}}.$$

Using the same covering number bound and the general result Corollary 23 (given in Section 3), we obtain the following fast rates result. A slightly more general result is easily derived by using Proposition 22 instead.

Theorem 8 *Let $\lambda > e/4$, $np \geq 20$ and $m \geq 5000$ with ν a distribution on $\mathbb{S}^{n-1}$. Then with probability at least $1 - e^{-x}$ over the m samples in E_m drawn according to ν, for all D with unit length columns:*

$$Eh_{R_\lambda,D} \leq 1.1 E_m h_{R_\lambda,D} + 9\frac{np\ln\left(4\lambda m\right) + x}{m}.$$

Generalization bounds for k sparse representation. Proposition 3 and Corollary 4 imply certain generalization bounds for the problem of dictionary learning for k sparse representation, which we give here.

A straight forward combination of Theorem 2 of Maurer and Pontil (2010) (given here as Theorem 6) and Proposition 3 results in the following theorem.

Theorem 9 *Let $\delta < 1$ with ν a distribution on $\mathbb{S}^{n-1}$. Then with probability at least $1-e^{-x}$ over the m samples in E_m drawn according to ν, for all D s.t. $\mu_{k-1}(D) \le \delta$ and with unit length columns:*

$$Eh^2_{H_k,D} \le E_m h^2_{H_k,D} + \frac{p}{\sqrt{m}}\left(\frac{14k}{1-\delta} + \frac{1}{2}\sqrt{\ln\left(16m\left(\frac{k}{1-\delta}\right)^2\right)}\right) + \sqrt{\frac{x}{2m}}.$$

In the case of clustering we have $k = 1$ and $\delta = 0$ and this result approaches the rates of Biau et al. (2008).

The following theorems follow from the covering number bound of Corollary 4 and applying the general results of Section 3 as for the l_1 sparsity results.

Theorem 10 *Let $\delta < 1$ with ν a distribution on $\mathbb{S}^{n-1}$. Then with probability at least $1-e^{-x}$ over the m samples in E_m drawn according to ν, for all D s.t. $\mu_{k-1}(D) \le \delta$ and with unit length columns:*

$$Eh_{H_k,D} \le E_m h_{H_k,D} + \sqrt{\frac{np\ln\frac{4\sqrt{m}k}{1-\delta}}{2m}} + \sqrt{\frac{x}{2m}} + \sqrt{\frac{4}{m}}.$$

Theorem 11 *Let $\delta < 1$, $np \ge 20$ and $m \ge 5000$ with ν a distribution on $\mathbb{S}^{n-1}$. Then with probability at least $1-e^{-x}$ over the m samples in E_m drawn according to ν, for all D s.t. $\mu_{k-1}(D) \le \delta$ and with unit length columns:*

$$Eh_{H_k,D} \le 1.1E_m h_{H_k,D} + 9\frac{np\ln\left(\frac{4\sqrt{m}k}{1-\delta}\right) + x}{m}.$$

Generalization bounds for dictionary learning in feature spaces. We further consider applications of dictionary learning to signals that are not represented as elements in a vector space, or that have a very high (possibly infinite) dimension.

In addition to providing an approximate reconstruction of signals, sparse representation can also be considered as a form of analysis, if we treat the choice of non zero coefficients and their magnitude as features of the signal. In the domain of images, this has been used to perform classification (in particular, face recognition) by Wright et al. (2008). Such analysis does not require that the data itself be represented in $\mathbb{R}^n$ (or in any vector space); it is enough that the similarity between data elements is induced from an inner product in a feature space. This requirement is fulfilled by using an appropriate kernel function.

Definition 12 *Let $\mathcal{R}$ be a set of data representations, and let the kernel function $\kappa : \mathcal{R}^2 \to \mathbb{R}$ and the feature mapping $\phi : \mathcal{R} \to \mathcal{H}$ be such that:*

$$\kappa\left(x, y\right) = \left\langle \phi\left(x\right), \phi\left(y\right)\right\rangle$$

where $\mathcal{H}$ is some Hilbert space.

As a concrete example, choose a sequence of n words, and let ϕ map a document to the vector of counts of appearances of each word in it (also called bag of words). Treating $\kappa(a,b) = \langle \phi(a), \phi(b) \rangle$ as the similarity between documents a and b, is the well known "bag of words" approach, applicable to many document related tasks (Shawe-Taylor and Cristianini, 2004). Then the statement $\phi(a) + \phi(b) \approx \phi(c)$ does not imply that c can be reconstructed from a and b, but we might consider it indicative of the content of c. The dictionary of elements used for representation could be decided via dictionary learning, and it is natural to choose the dictionary so that the bags of words of documents are approximated well by small linear combinations of those in the dictionary.

As the example above suggests, the kernel dictionary learning problem is to find a dictionary D minimizing

$$\mathbb{E}_{x\sim\nu} h_{\phi,A,D}(x),$$

where we consider the representation error function

$$h_{\phi,A,D}(x) = \min_{a\in A} \left\| (\Phi D)\, a - \phi\,(x) \right\|_{\mathcal{H}},$$

in which Φ acts as ϕ on the elements of D, $A \in \{R_\lambda, H_k\}$, and the norm $\|\cdot\|_{\mathcal{H}}$ is that induced by the kernel on the feature space $\mathcal{H}$.

Analogues of all the generalization bounds mentioned so far can be replicated in the kernel setting. The dimension free results of Maurer and Pontil (2010) apply most naturally in this setting, and may be combined with our results to cover also dictionaries for k sparse representation, under reasonable assumptions on the kernel.

Proposition 13 *Let ν be any distribution on $\mathcal{R}$ such that $x \sim \nu$ implies that $\phi(x)$ is in the unit ball $B_{\mathcal{H}}$ of $\mathcal{H}$ with probability 1. Then with probability at least $1 - e^{-x}$ over the m samples in E_m drawn according to ν, for all $D \subset \mathcal{R}$ with cardinality p such that $\Phi D \subset B_{\mathcal{H}}$ and $\mu_{k-1}(\Phi D) \le \delta < 1$:*

$$Eh^2_{\phi,H_k,D} \le E_m h^2_{\phi,H_k,D} + \sqrt{\frac{p^2\left(14k/(1-\delta) + 1/2\sqrt{\ln\left(16m\left(\frac{k}{1-\delta}\right)^2\right)}\right)^2}{m}} + \sqrt{\frac{x}{2m}}.$$

Note that in $\mu_{k-1}(\Phi D)$ the Babel function is defined in terms of inner products in $\mathcal{H}$, and can therefore be computed efficiently by applications of the kernel.

In Section 5 we prove the above result and also cover number bounds as in the linear case considered before. In the current setting, these bounds depend on the Hölder smoothness order α of the feature mapping ϕ. Formal definitions are given in Section 5 but as an example, the well known Gaussian kernel has $\alpha = 1$. We give now one of the generalization bounds using this method.

Theorem 14 *Let $\mathcal{R}$ have ε covers of order $(C/\varepsilon)^n$. Let $\kappa : \mathcal{R}^2 \to \mathbb{R}^+$ be a kernel function s.t. $\kappa(x,y) = \langle \phi(X), \phi(Y) \rangle$, for ϕ which is uniformly L-Hölder of order $\alpha > 0$ over $\mathcal{R}$, and let $\gamma = \max_{x\in\mathcal{R}} \|\phi(x)\|_{\mathcal{H}}$. Let $\delta < 1$, and ν any distribution on $\mathcal{R}$, then with probability at* *least $1 - e^{-x}$ over the m samples in E_m drawn according to ν, for all dictionaries $D \subset \mathcal{R}$ of cardinality p s.t. $\mu_{k-1}(\Phi D) \leq \delta < 1$ (where Φ acts like ϕ on columns):*

$$Eh_{H_k,D} \leq E_m h_{H_k,D} + \gamma \left(\sqrt{\frac{np \ln\left(\sqrt{m} C^{\alpha} \frac{k\gamma^2 L}{1-\delta}\right)}{2\alpha m}} + \sqrt{\frac{x}{2m}} \right) + \sqrt{\frac{4}{m}}.$$

The covering number bounds needed to prove this theorem and analogs for the other generalization bounds are proved in Section 5.

3. Covering numbers of $\mathcal{G}_\lambda$ and $\mathcal{F}_{\delta,k}$

The main content of this section is the proof of Theorem 1 and Corollary 4. We also show that in the k sparse representation setting a finite bound on λ does not occur generally thus an additional restriction, such as the near-orthogonality on the set of dictionaries on which we rely in this setting, is necessary. Lastly, we recall known results from statistical learning theory that link covering numbers to generalization bounds.

We recall the definition of the covering numbers we wish to bound. Anthony and Bartlett (1999) give a textbook introduction to covering numbers and their application to generalization bounds.

Definition 15 (Covering number) *Let (M, d) be a metric space and $S \subset M$. Then the ε covering number of S defined as $N(\varepsilon, S, d) = \min\left\{|A| \,|A \subset M \text{ and } S \subset \left(\bigcup_{a\in A} B_d(a, \varepsilon)\right)\right\}$ is the size of the minimal ε cover of S using d.*

To prove Theorem 1 and Corollary 4 we first note that the space of all possible dictionaries is a subset of a unit ball in a Banach space of dimension np (with a norm specified below). Thus (see formalization in Proposition 5 of Cucker and Smale, 2002) the space of dictionaries has an ε cover of size $(4/\varepsilon)^{np}$. We also note that a uniformly L Lipschitz mapping between metric spaces converts ε/L covers into ε covers. Then it is enough to show that Ψ_λ defined as $D \mapsto h_{R_\lambda,D}$ and Φ_k defined as $D \mapsto h_{H_k,D}$ are uniformly Lipschitz (when Φ_k is restricted to the dictionaries with $\mu_{k-1}(D) \leq c < 1$). The proof of these Lipschitz properties is our next goal, in the form of Lemmas 18 and 19.

The first step is to be clear about the metrics we consider over the spaces of dictionaries and of error functions.

Definition 16 (Induced matrix norm) *Let $p, q \geq 1$, then a matrix $A \in \mathbb{R}^{n\times m}$ can be considered as an operator $A : \left(\mathbb{R}^m, \|\cdot\|_p\right) \to \left(\mathbb{R}^n, \|\cdot\|_q\right)$. The p, q induced norm is $\|A\|_{p,q} \triangleq \sup_{x\in\mathbb{R}^m \|x\|_p=1} \|Ax\|_q$.*

Lemma 17 *For any matrix D, $\|D\|_{1,2}$ is equal to the maximal Euclidean norm of any column in D.*

Proof That the maximal norm of a column bounds $\|D\|_{1,2}$ can be seen geometrically; $Da/\|a\|_1$ is a convex combination of column vectors, then $\|Da\|_2 \leq \max_{d_i} \|d_i\|_2 \|a\|_1$ because a norm is convex. Equality is achieved for $a = e_i$, where d_i is the column of maximal norm. ■

The images of Ψ_λ and Φ_k are sets of representation error functions–each dictionary induces a set of precisely representable signals, and a representation error function is simply a map of distances from this set. Representation error functions are clearly continuous, 1-Lipschitz, and into $[0,1]$. In this setting, a natural norm over the images is the supremum norm $\|\cdot\|_\infty$.

Lemma 18 *The function Ψ_λ is λ-Lipschitz from $\left(\mathbb{R}^{n\times m}, \|\cdot\|_{1,2}\right)$ to $C\left(\mathbb{S}^{n-1}\right)$.*

Proof Let D and D' be two dictionaries whose corresponding elements are at most $\varepsilon > 0$ far from one another. Let x be a unit signal and Da an optimal representation for it. Then $\|(D-D')\,a\|_2 \leq \|D-D'\|_{1,2}\,\|a\|_1 \leq \varepsilon\lambda$. If $D'a$ is very close to Da in particular it is not a much worse representation of x, and replacing it with the optimal representation under D', we have $h_{R_\lambda,D'}(x) \leq h_{R_\lambda,D}(x) + \varepsilon\lambda$. By symmetry we have $|\Psi_\lambda(D)(x) - \Psi_\lambda(D')(x)| \leq \lambda\varepsilon$. This holds for all unit signals, then $\|\Psi_\lambda(D) - \Psi_\lambda(D')\|_\infty \leq \lambda\varepsilon$. ■

We now provide a proof for Proposition 3 which is used in the corresponding treatment for covering numbers under k sparsity.

Proof (Of Proposition 3) Let D^k be a submatrix of D whose k columns from D achieve the minimum on $h_{H_k,D}(x)$ for $x \in \mathbb{S}^{n-1}$. We now consider the Gram matrix $G = \left(D^k\right)^\top D^k$ whose diagonal entries are the norms of the elements of D^k, therefore at least 1. By the Gersgorin theorem (Horn and Johnson, 1990), each eigenvalue of a square matrix is "close" to a diagonal entry of the matrix; the absolute difference between an eigenvalue and its diagonal entry is upper bounded by the sum of the absolute values of the remaining entries of the same row. Since a row in G corresponds to the inner products of an element from D^k with every element from D^k, this sum is upper bounded by δ for all rows. Then we conclude the eigenvalues of the Gram matrix are lower bounded by $1-\delta > 0$. Then in particular G has a symmetric inverse G^{-1} whose eigenvalues are positive and bounded from above by $1/\left(1-\delta\right)$. The maximal magnitude of an eigenvalue of a symmetric matrix coincides with its induced norm $\|\cdot\|_{2,2}$, therefore $\left\|G^{-1}\right\|_{2,2} \leq 1/(1-\delta)$.

Linear dependence of elements of D^k would imply a non-trivial nullspace for the invertible G. Then the elements of D^k are linearly independent, which implies that the unique optimal representation of x as a linear combination of the columns of D^k is $D^k a$ with

$$a = \left(\left(D^k\right)^\top D^k\right)^{-1}\left(D^k\right)^\top x.$$

Using the above and the definition of induced matrix norms, we have

$$\|a\|_2 \leq \left\|\left(\left(D^k\right)^\top D^k\right)^{-1}\right\|_{2,2}\left\|\left(D^k\right)^\top x\right\|_2 \leq 1/(1-\delta)\left\|\left(D^k\right)^\top x\right\|_2.$$

The vector $\left(D^k\right)^\top x$ is in $\mathbb{R}^k$ and by the Cauchy Schwartz inequality $\langle d_i, x\rangle \leq \gamma$, then $\left\|\left(D^k\right)^\top x\right\|_2 \leq \sqrt{k}\left\|\left(D^k\right)^\top x\right\|_\infty \leq \sqrt{k}\gamma$. Since only k entries of a are non zero, $\|a\|_1 \leq$

$\sqrt{k}\left\|a\right\|_2 \leq k\gamma/(1-\delta)$. ■

Lemma 19 *The function Φ_k is a $k/(1-\delta)$-Lipschitz mapping from the set of normalized dictionaries with $\mu_{k-1}(D) < \delta$ with the metric induced by $\left\|\cdot\right\|_{1,2}$ to $C\left(\mathbb{S}^{n-1}\right)$.*

The proof of this lemma is the same as that of Lemma 18, except that a is taken to be an optimal representation that fulfills $\left\|a\right\|_1 \leq \lambda = k/\left(1-\mu_{k-1}(D)\right)$, whose existence is guaranteed by Proposition 3.

This concludes the proof of Theorem 1 and Corollary 4.

The next theorem shows that unfortunately, Φ is *not* uniformly L-Lipschitz for any constant L, requiring its restriction to an appropriate subset of the dictionaries.

Theorem 20 *For any $1 < k < n, p$, there exists $c > 0$ and q, such that for every $\varepsilon > 0$, there exist D, D' such that $\left\|D - D'\right\|_{1,2} < \varepsilon$ but $\left|\left(h_{H_k,D}(q) - h_{H_k,D'}(q)\right)\right| > c$.*

Proof First we show that for any dictionary D there exist $c > 0$ and $x \in \mathbb{S}^{n-1}$ such that $h_{H_k,D}(x) > c$. Let $\nu_{\mathbb{S}^{n-1}}$ be the uniform probability measure on the sphere, and A_c the probability assigned by it to the set within c of a k dimensional subspace. As $c \searrow 0$, A_c also tends to zero, then there exists $c > 0$ s.t. $\binom{p}{k}A_c < 1$. Then for that c and any dictionary D there exists a set of positive measure on which $h_{H_k,D} > c$, let q be a point in this set. Since $h_{H_k,D}(x) = h_{H_k,D}(-x)$, we may assume without loss of generality that $\langle e_1, q\rangle \geq 0$.

We now fix the dictionary D; its first $k-1$ elements are the standard basis $\{e_1, \ldots, e_{k-1}\}$, its kth element is $D_k = \sqrt{1-\varepsilon^2/4}e_1 + \varepsilon e_k/2$, and the remaining elements are chosen arbitrarily. Now construct D' to be identical to D except its kth element is $v = \sqrt{1-\varepsilon^2/4}e_1 + lq$ choosing l so that $\left\|v\right\|_2 = 1$. Then there exist $a, b \in \mathbb{R}$ such that $q = aD'_1 + bD'_k$ and we have $h_{H_k,D'}(q) = 0$, fulfilling the second part of the theorem. On the other hand, since $\langle e_1, q\rangle \geq 0$, we have $l \leq \varepsilon/2$, and then we find $\left\|D - D'\right\|_{1,2} = \left\|\varepsilon e_k/2 - lq\right\|_2 \leq \left\|\varepsilon e_k/2\right\| + \left\|lq\right\| = \varepsilon/2 + l \leq \varepsilon$. ■

To conclude the generalization bounds of Theorems 7, 8, 10, 11 and 14 from the covering number bounds we have provided, we use the following results. The first result is a straight forward application of Hoeffding's inequality, a union bound and the l_∞ cover number bounds. The second result[1] (along with its corollary) gives fast rate bounds and uses the $\left\|\cdot\right\|_\infty$ cover number bounds to achieve better constants for this problem than the more general results by Mendelson (2003) and Bartlett et al. (2005).

Lemma 21 *Let $\mathcal{F}$ be a class of $[0, B]$ functions with covering number bound $(C/\varepsilon)^d > e/B^2$ under the supremum norm. Then for every $x > 0$, with probability of at least $1 - e^{-x}$ over the m samples in E_m chosen according to ν, for all $f \in \mathcal{F}$:*

$$Ef \leq E_m f + B\left(\sqrt{\frac{d\ln\left(C\sqrt{m}\right)}{2m}} + \sqrt{\frac{x}{2m}}\right) + \sqrt{\frac{4}{m}}.$$

[1] We thank Andreas Maurer for suggesting this result and a proof elaborated in the full version.

Proposition 22 *Let $\mathcal{F}$ be a class of $[0,1]$ functions that can be covered for any $\varepsilon > 0$ by at most $(C/\varepsilon)^d$ balls of radius ε in the L_∞ metric where $C \geq e$ and $\beta > 0$. Then with probability at least $1 - \exp(-x)$, we have for all $f \in \mathcal{F}$:*

$$Ef \leq (1+\beta) E_m f + K(d,m,\beta) \frac{d \ln(Cm) + x}{m},$$

where $K(d,m,\beta) = \sqrt{2\left(\frac{9}{\sqrt{m}} + 2\right)\left(\frac{d+3}{3d}\right) + 1} + \left(\frac{9}{\sqrt{m}} + 2\right)\left(\frac{d+3}{3d}\right) + 1 + \frac{1}{2\beta}$.

The corollary we use to obtain Theorems 8 and 11 follows because $K(d,m,\beta)$ is non-increasing in d, m.

Corollary 23 *Let $\mathcal{F}, x$ be as above. For $d \geq 20$, $m \geq 5000$ and $\beta = 0.1$ we have with probability at least $1 - \exp(-x)$ for all $f \in \mathcal{F}$:*

$$Ef \leq 1.1 E_m f + 9 \frac{d \ln(Cm) + x}{m}.$$

4. On the Babel function

The Babel function is one of several metrics defined in the sparse representations literature to quantify an "almost orthogonality" property that dictionaries may enjoy. Such properties have been shown to imply theoretical properties such as uniqueness of the optimal k sparse representation. In the algorithmic context, Donoho and Elad (2003) and Tropp (2004) use the Babel function to show that particular efficient algorithms for finding sparse representations fulfill certain quality guarantees when applied to such dictionaries. This reinforces the practical importance of the learnability of this class of dictionary. We proceed to discuss some elementary properties of the Babel function, and then state a bound on the proportion of dictionaries having sufficiently good Babel function.

Measures of orthogonality are typically defined in terms of inner products between the elements of the dictionary. Perhaps the simplest of these measures of orthogonality is the following special case of the Babel function.

Definition 24 *The coherence of a dictionary D is $\mu_1(D) = \max_{i \neq j} |\langle d_i, d_j \rangle|$.*

The Babel function considers sums of k inner products at a time rather than the maximum over all inner products, and thus better quantifies the effects of non orthogonality on representing a signal with particular level $k+1$ of sparsity. As a particular example of the finer grained control μ_k when compared to μ_1, consider the following example. Let D consist of k pairs of elements, so that the subspace spanned by each pair is orthogonal to all other elements, and such that the inner product between the elements of any single pair is half. In this case $\mu_k(D) = \mu_1(D) = 1/2$. However note that to ensure $\mu_k < 1$ only restricting μ_1 requires the constraint $\mu_1(D) < 1/k$, which is not fulfilled in our example.

To better understand $\mu_k(D)$, we consider first its extreme values. When $\mu_k(D) = 0$, for any $k > 1$, this means that D is an orthogonal set (therefore $p \leq n$). The maximal value of $\mu_k(D)$ is k, and occurs only if some dictionary element is repeated (up to sign) at least $k+1$ times.

A well known generic class of dictionaries with more elements than a basis is that of *frames* (see Duffin and Schaeer, 1952), which include many wavelet systems and filter banks. Some frames can be trivially seen to fulfill our condition on the Babel function.

Proposition 25 *Let $D \in \mathbb{R}^{n\times p}$ be a frame of $\mathbb{R}^n$, so that for every $v \in \mathbb{S}^{n-1}$ we have that $\sum_{i=1}^{n} |\langle v, d_i\rangle|^2 \leq B$, with $\|d_i\|_2 = 1$ for all i, and $B < 1 + 1/k$. Then $\mu_{k-1}(D) < 1$.*

This may be easily verified by considering the inner products of any dictionary element with any other k elements as a vector in $\mathbb{R}^k$; the frame condition bounds its squared Euclidean norm by $B-1$ (we remove the inner product of the element with itself in the frame expression). Then use the equivalence of l_1 and l_2 norms.

4.1. Proportion of dictionaries with $\mu_{k-1}(D) < \delta$

We return to the question of the prevalence of dictionaries having $\mu_{k-1} < \delta$. Are almost all dictionaries such? If the answer is affirmative, it implies that Theorem 11 is quite strong, and representation finding algorithms such as basis pursuit are almost always exact, which might help prove properties of dictionary learning algorithms. If the opposite is true and few dictionaries have low Babel function, the results of this paper are weak. While there might be better probability measures on the space of dictionaries, we consider one that seems natural: suppose that a dictionary D is constructed by choosing p unit vectors uniformly from $\mathbb{S}^{n-1}$; what is the probability that $\mu_{k-1}(D) < \delta$?

Theorem 5 gives us the following answer to this question. Under the assumption that the sparsity parameter k grows slowly, if at all, as $n \nearrow \infty$ (specifically, that $k \ln p = o(\sqrt{n})$), this theorem implies that asymptotically *almost all dictionaries under the Lebesgue measure are learnable.*

5. Dictionary learning in feature spaces

We propose in Section 2 a scenario in which dictionary learning is performed in a feature space corresponding to a kernel function. Here we show how to adapt the different generalization bounds discussed in this paper for the particular case of $\mathbb{R}^n$ to more general feature spaces, and the dependence of the sample complexities on the properties of the kernel function or the corresponding feature mapping. We begin with the relevant specialization of the results of Maurer and Pontil (2010) which have the simplest dependence on the kernel, and then discuss the extensions to k sparse representation and to the cover number techniques presented in the current work.

Theorem 6 applies as is to the feature space, under the simple assumption that the dictionary elements and signals are in its unit ball which is guaranteed by some kernels such as the Gaussian kernel. Then we take ν on the unit ball of $\mathcal{H}$ to be induced by some distribution ν' on the domain of the kernel, and the theorem applies to any such ν' on $\mathcal{R}$. Nothing more is required if the representation is chosen from R_λ. The corresponding generalization bound for k sparse representations when the dictionary elements are near orthogonal in the feature space is given in Proposition 13.

Proof (Of Proposition 13) Proposition 3 applies with the Euclidean norm of $\mathcal{H}$, and $\gamma = 1$. We apply Theorem 6 with $\lambda = k/(1-\delta)$. ■

The results so far show that generalization in dictionary learning can occur despite the potentially infinite dimension of the feature space, without considering practical issues of representation and computation. We now make the domain and applications of the kernel explicit in order to address a basic computational question, and allow the use of cover number based generalization bounds to prove Theorem 14. We now consider signals represented in a metric space $(\mathcal{R}, d)$, in which similarity is measured by the kernel κ corresponding to the feature map $\phi : \mathcal{R} \to \mathcal{H}$. The elements of a dictionary D are now from $\mathcal{R}$, and we denote ΦD their mapping by ϕ to $\mathcal{H}$. The representation error function used is $h_{\phi,A,D}$.

We now show that the approximation error in the feature space is a quadratic function of the coefficient vector; the quadratic function for particular D and x may be found by applications of the kernel.

Proposition 26 *Computing the representation error at a given x, a, D requires $O\left(p^2\right)$ kernel applications in general, and only $O\left(k^2 + p\right)$ when a is k sparse.*

The squared error expands to

$$\sum_{i=1}^{p} a_i \sum_{j=1}^{p} a_j \kappa\left(d_i, d_j\right) + \kappa\left(x, x\right) - 2 \sum_{i=1}^{p} a_i \kappa\left(x, d_i\right).$$

We note that the k sparsity constraint on a poses algorithmic difficulties beyond those addressed here. Some of the common approaches to these, such as orthogonal matching pursuit (Chen et al., 1989), also depend on the data only through their inner products, and may therefore be adapted to the kernel setting.

The cover number bounds depend strongly on the dimension of the space of dictionary elements. Taking $\mathcal{H}$ as the space of dictionary elements is the simplest approach, but may lead to vacuous or weak bounds, for example in the case of the Gaussian kernel whose feature space is infinite dimensional. Instead we propose to use the space of data representations $\mathcal{R}$, whose dimensions are generally bounded by practical considerations. In addition, we will assume that the kernel is not "too wild" in the following sense.

Definition 27 *Let $L, \alpha > 0$, and let (A, d') and (B, d) be metric spaces. We say a mapping $f : A \to B$ is uniformly L Hölder of order α on a set $S \subset A$ if $\forall x, y \in S$, the following bound holds:*

$$d\left(f(x), f(y)\right) \leq L \cdot d'(x, y)^{\alpha}.$$

The relevance of this smoothness condition is as follows.

Lemma 28 *A Hölder function maps an ε cover of S to an $L\varepsilon^{\alpha}$ cover of its image $f(S)$. Thus, to obtain an ε cover of the image of S, it is enough to begin with an $\left(\varepsilon/L\right)^{1/\alpha}$ cover of S.*

A Hölder feature map ϕ allows us to bound the cover numbers of the dictionary elements in $\mathcal{H}$ using their cover number bounds in $\mathcal{R}$. Note that not every kernel corresponds to a Hölder feature map (the Dirac δ kernel is a counter example: any two distinct elements are mapped to elements at a mutual distance of 1), and sometimes analyzing the feature map is harder than analyzing the kernel. The following lemma bounds the geometry of the feature map using that of the kernel.

Lemma 29 *Let $\kappa(x,y) = \langle \phi(x), \phi(y) \rangle$, and assume further that κ fulfills a Hölder condition of order α uniformly in each parameter, that is, $|\kappa(x,y) - \kappa(x+h,y)| \le L \|h\|^{\alpha}$. Then ϕ uniformly fulfills a Hölder condition of order $\alpha/2$ with constant $\sqrt{2L}$.*

This result is not sharp. For example, for the Gaussian case, both kernel and the feature map are Hölder order 1.

Proof Using the Hölder condition, we have that $\|\phi(x) - \phi(y)\|_{\mathcal{H}}^2 = \kappa(x,x) - \kappa(x,y) + \kappa(y,y) - \kappa(x,y) \le 2L \|x-y\|^{\alpha}$. All that remains is to take the square root of both sides. ∎

For a given feature mapping ϕ, set of representations $\mathcal{R}$, we define two families of function classes so:

$$\begin{aligned} \mathcal{W}_{\phi,\lambda} &= \left\{ h_{\phi,R_\lambda,D} : D \in \mathcal{D}^p \right\} \text{and} \\ \mathcal{Q}_{\phi,k,\delta} &= \left\{ h_{\phi,H_k,D} : D \in \mathcal{D}^p \wedge \mu_{k-1}(\Phi D) \le \delta \right\}. \end{aligned}$$

The next proposition completes this section by giving the cover number bounds for the representation error function classes induced by appropriate kernels, from which various generalization bounds easily follow, such as Theorem 14.

Proposition 30 *Let $\mathcal{R}$ be a set of representations with a cover number bound of $(C/\varepsilon)^n$, and let either ϕ be uniformly L Hölder condition of order α on $\mathcal{R}$, or κ be uniformly L Hölder of order 2α on $\mathcal{R}$ in each parameter, and let $\gamma = \sup_{d \in \mathcal{R}} \|\phi(d)\|_{\mathcal{H}}$. Then the function classes $\mathcal{W}_{\phi,\lambda}$ and $\mathcal{Q}_{\phi,k,\delta}$ taken as metric spaces with the supremum norm, have ε covers of cardinalities at most $\left(C\left(\lambda\gamma L/\varepsilon\right)^{1/\alpha}\right)^{np}$ and $\left(C\left(k\gamma^2 L/\left(\varepsilon\left(1-\delta\right)\right)\right)^{1/\alpha}\right)^{np}$, respectively.*

Proof We first consider the case of l_1 constrained coefficients. If $\|a\|_1 \le \lambda$ and also $\max_{d \in \mathcal{D}} \|\phi(d)\|_{\mathcal{H}} \le \gamma$ then by considerations applied in Section 3, to obtain an ε cover of the set $\left\{\min_a \|(\Phi D)\, a - \phi(x)\|_{\mathcal{H}} : D \in \mathcal{D}\right\}$, it is enough to obtain an $\varepsilon/(\lambda\gamma)$ cover of $\{\Phi D : D \in \mathcal{D}\}$. If also ϕ is uniformly L Hölder of order α over $\mathcal{R}$ then an $(\lambda\gamma L/\varepsilon)^{-1/\alpha}$ cover of the set of dictionaries is sufficient, which as we have seen requires at most $\left(C\left(\lambda\gamma L/\varepsilon\right)^{1/\alpha}\right)^{np}$ elements.

In the case of l_0 constrained representation, the bound on λ due to Proposition 3 is $\gamma k(1-\delta)$, and the result follows from the above by substitution. ∎

6. Conclusions

Our work has several implications on the design of dictionary learning algorithms as used in signal, image, and natural language processing. First, the fact that generalization is only logarithmically dependent on the l_1 norm of the coefficient vector widens the set of applicable approaches to penalization. Second, in the particular case of k sparse representation, we have shown that the Babel function is a key property for the generalization of dictionaries. It might thus be useful to modify dictionary learning algorithms so that they obtain dictionaries with low Babel functions, possibly through regularization or through certain convex relaxations. Third, mistake bounds (e.g., Mairal et al. 2010) on the quality of the solution to the coefficient finding optimization problem may lead to generalization bounds for practical algorithms, by tying such algorithms to k sparse representation.

The upper bounds presented here invite complementary lower bounds. The existing lower bounds for $k = 1$ (vector quantization) and for $k = p$ (representation using PCA directions) are applicable, but do not capture the geometry of general k sparse representation, and in particular do not clarify the effective dimension of the unrestricted class of dictionaries for it. We have not excluded the possibility that the class of unrestricted dictionaries has the same dimension as that of those with a small Babel function. The best upper bound we know for the larger class, being the trivial one of order $O\left(\binom{p}{k}n^2 / m\right)$, leaves a significant gap for future exploration.

We view the dependence on μ_{k-1} from an "algorithmic luckiness" perspective (Herbrich and Williamson, 2003): if the data is described by a dictionary with low Babel function the generalization bounds are encouraging.

Acknowledgments

We thank Shahar Mendelson for helpful discussions. A.B. was partly supported by the European Communitys FP7-FET program, SMALL project, under grant agreement no. 225913. S.M and D.V. were partially supported by the ISF under contract 890015.

References

Michal Aharon, Michael Elad, and Alfred M. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations.* Cambridge University Press, 1999.

Peter L. Bartlett, Tamás Linder, and Gábor Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information theory*, 44(5):1802–1813, 1998.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33:1497–1537, 2005.

Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294, 2007.

Alfred M. Bruckstein, David L. Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

Emmanuel J. Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12): 5406–5425, 2006.

Scott S. Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

Sheng Chen, Stephen A. Billings, and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50 (5):1873–1896, 1989.

Felipe Cucker and Stephen Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc*, 39(1):1–50, 2002.

Geoff Davis, Stèphane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

Richard J. Duffin and Albert C. Schaeer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc*, 72:341–366, 1952.

Ralf Herbrich and Robert Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3:175–212, 2003.

Roger A. Horn and Charles R. Johnson. *Matrix analysis.* Cambridge University Press, 1990.

Kenneth Kreutz-Delgado, Joseph F. Murray, Bhaskar D. Rao, Kjersti Engan, Te-Won Lee, and Terrance J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems 19*, pages 801–808. MIT Press, Cambridge, MA, 2007.

Michael S. Lewicki, Terrence J. Sejnowski, and Howard Hughes. Learning overcomplete representations. *Neural Computation*, 12:337–365, 1998.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

Andreas Maurer and Massimiliano Pontil. K-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56:5839–5846, November 2010.

Shahar Mendelson. A few notes on statistical learning theory. *Advanced Lectures on Machine Learning*, pages 1–40, 2003.

Bruno A. Olshausen and David J. Fieldt. Sparse coding with an overcomplete basis set: a strategy employed by V1. *Vision Research*, 37:3311–3325, 1997.

Gabriel Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

Matan Protter and Michael Elad. Sparse and redundant representations and motion-estimation-free algorithm for video denoising. *Wavelets XII. Proceedings of the SPIE*, 6701:43, 2007.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis.* Cambridge University Press, 2004.

John Shawe-Taylor, Christopher K. I. Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005.

Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50:2231–2242, 2004.

John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 210–227, 2008.