Testing Sparsity over Known and Unknown Bases

Siddharth Barman [1] **Arnab Bhattacharyya** [1] **Suprovat Ghoshal** [1]

Abstract

Sparsity is a basic property of real vectors that is exploited in a wide variety of machine learning applications. In this work, we describe *property testing* algorithms for sparsity that observe a low-dimensional projection of the input. We consider two settings. In the first setting, we test sparsity with respect to an unknown basis: given input vectors $\mathbf{y}_1, \ldots, \mathbf{y}_p \in \mathbb{R}^d$ whose concatenation as columns forms $\mathbf{Y} \in \mathbb{R}^{d\times p}$, does $\mathbf{Y} = \mathbf{AX}$ for matrices $\mathbf{A} \in \mathbb{R}^{d\times m}$ and $\mathbf{X} \in \mathbb{R}^{m\times p}$ such that each column of $\mathbf{X}$ is k-sparse, or is $\mathbf{Y}$ "far" from having such a decomposition? In the second setting, we test sparsity with respect to a known basis: for a fixed design matrix $\mathbf{A} \in \mathbb{R}^{d\times m}$, given input vector $\mathbf{y} \in \mathbb{R}^d$, is $\mathbf{y} = \mathbf{Ax}$ for some k-sparse vector $\mathbf{x}$ or is $\mathbf{y}$ "far" from having such a decomposition? We analyze our algorithms using tools from high-dimensional geometry and probability.

1. Introduction

Property testing is the study of algorithms that query their input a small number of times and distinguish between whether their input satisfies a given property or is "far" from satisfying that property. The quest for efficient testing algorithms was initiated by (Blum et al., 1993) and (Babai et al., 1991) and later explicitly formulated by (Rubinfeld & Sudan, 1996) and (Goldreich et al., 1998). Property testing can be viewed as a relaxation of the traditional notion of a decision problem, where the relaxation is quantified in terms of a distance parameter. There has been extensive work in this area over the last couple of decades; see, for instance, the surveys (Ron, 2008) and (Rubinfeld & Shapira, 2006) for some different perspectives.

As evident from these surveys, research in property testing has largely focused on properties of combinatorial and algebraic structures, such as bipartiteness of graphs, linearity of Boolean functions on the hypercube, membership in error-correcting codes or representability of functions as concise Boolean formulae. In this work, we study the question of testing properties of *continuous* structures, specifically properties of vectors and matrices over the reals.

Our computational model is a natural extension of the standard property testing framework by allowing queries to be linear measurements of the input. Let $\mathcal{P} \subset \mathbb{R}^d$ be a property of real vectors. Let $\mathsf{dist} : \mathbb{R}^d \to \mathbb{R}^{\geqslant 0}$ be a "distance" function such that $\mathsf{dist}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{P}$. We say that an algorithm $\mathcal{A}$ is a *tester for* $\mathcal{P}$ with respect to dist and with parameters $\varepsilon, \delta > 0$ if for any input $\mathbf{y} \in \mathbb{R}^n$, the algorithm $\mathcal{A}$ observes $\mathbf{My}$ where $\mathbf{M} \in \mathbb{R}^{q\times d}$ is a randomized matrix and has the following guarantee:

(i) If $\mathbf{y} \in \mathcal{P}$, $\mathbf{Pr}_{\mathbf{M}}[\mathcal{A}(\mathbf{My}) \text{ accepts}] \geqslant 1 - \delta$.

(ii) If $\mathsf{dist}(\mathbf{y}) > \varepsilon$, $\mathbf{Pr}_{\mathbf{M}}[\mathcal{A}(\mathbf{My}) \text{ accepts}] \leqslant \delta$.

We call each inner product between the rows of $\mathbf{M}$ and $\mathbf{y}$ a *(linear) query*, and the number of rows $q = q(\varepsilon, \delta)$ is the *query complexity* of the tester. The *running time* of the tester $\mathcal{A}$ is its running time on the outcome of its queries. As typical in property testing, we do not count the time needed to evaluate the queries. If $\mathcal{P} \subset \mathbb{R}^{d\times p}$ is a property of real matrices with an associated distance function $\mathsf{dist} : \mathbb{R}^{d\times p} \to \mathbb{R}^{\geqslant 0}$, testing is defined similarly: given an input matrix $\mathbf{Y} \in \mathbb{R}^{d\times p}$, the algorithm observes $\mathbf{MY}$ for a random matrix $\mathbf{M} \in \mathbb{R}^{q\times d}$ with analogous completeness and soundness properties. A linear projection of an input vector or matrix to a low-dimensional space is also called a *linear sketch* or a *linear measurement*. The technique of obtaining small linear sketches of high-dimensional vectors has been used to great effect in algorithms for streaming (e.g., (Alon et al., 1996; McGregor, 2014)) and numerical linear algebra (see (Woodruff, 2014) for an excellent survey). Because GPUs are specially designed to optimize matrix-vector computation, many modern optimization and learning algorithms work with linear sketches of their input.

We focus on testing whether a vector is **sparse** with respect

[*] Equal contribution [1] Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India. Correspondence to: Siddharth Barman <barman@iisc.ac.in>, Arnab Bhattacharyya <arnabb@iisc.ac.in>, Suprovat Ghoshal <suprovat@iisc.ac.in>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

to some basis.[1] A vector $\mathbf{x}$ is said to be *k-sparse* if it has at most k nonzero coordinates. Sparsity is a structural characteristic of signals of interest in a diverse range of applications. It is a pervasive concept throughout modern statistics and machine learning, and algorithms to solve inverse problems under sparsity constraints are among the most successful stories of the optimization community (see the book (Hastie et al., 2015)). The natural property testing question we consider is whether there exists a solution to a linear inverse problem under a sparsity constraint.

There are two settings in which we investigate the sparsity testing problem.

(a) In the first setting, the basis is not known in advance. For input vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p \in \mathbb{R}^d$, the property to test is whether there exists a matrix $\mathbf{A} \in \mathbb{R}^{d\times m}$ and k-sparse unit vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_p \in \mathbb{R}^m$ such that $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$ for all $i \in [p]$. Note that m is specified as a parameter and could be much larger than d (the *overcomplete* case). In this setting, we restrict the unknown $\mathbf{A}$ to be a (ε, k)-*RIP* matrix which means that $(1-\varepsilon)\|\mathbf{x}\| \leqslant \|\mathbf{A}\mathbf{x}\| \leqslant (1+\varepsilon)\|\mathbf{x}\|$ for any k-sparse $\mathbf{x}$. This is a standard assumption made in many related works (see Section 1.2 for details).

In this setting, we design an efficient tester for this property that projects the inputs to $O(\varepsilon^{-2}\log p)$ dimensions and, informally speaking, rejects if for all (ε, k)-RIP matrices $\mathbf{A}$, there is some $\mathbf{y}_i$ such that $\mathbf{y}_i - \mathbf{A}\mathbf{x}_i$ has large norm for all "approximately sparse" $\mathbf{x}_i$.

(b) In the second setting, a design matrix $\mathbf{A} \in \mathbb{R}^{d\times m}$ is known explicitly, and the property to test is whether a given input vector $\mathbf{y} \in \mathbb{R}^d$ equals $\mathbf{A}\mathbf{x}$ for a k-sparse vector $\mathbf{x} \in \mathbb{R}^m$. For instance, $\mathbf{A}$ can be the Fourier basis or an overcomplete dictionary in an image processing application. We approach this problem in full generality, without putting any restriction on the structure of $\mathbf{A}$.

Informally, our main result in this setting is that for any design matrix $\mathbf{A}$, there exists a tester projecting the input $\mathbf{y}$ to $O(k\log m)$ dimensions that rejects if $\mathbf{y}-\mathbf{A}\mathbf{x}$ has large norm for any $O(k)$-sparse $\mathbf{x}$. The running time of the tester is polynomial in m. As we describe in Section 1.2, previous work in numerical linear algebra yields a tester with the same query complexity and with qualitatively similar soundness guarantees but which requires running time *exponential* in m or assumptions about the matrix $\mathbf{A}$.

Remark 1.1 (Problem Formulation). *Note that the settings considered in the known and unknown design matrix settings are quite different from each other. In particular, for the known design setting, the input is a single vector. However, given a single input vector $\mathbf{y} \in \mathbb{R}^d$, the analogous unknown design testing question would be moot, since one can always consider the vector $\mathbf{y}$ to be the design matrix $\mathbf{A}$, in which it trivially admits a 1-sparse representation. For the same reason, unknown design testing is interesting only when the number of vectors p exceeds m.*

In both of the above tests, the measurement matrix is a random matrix with iid gaussian entries, chosen so as to preserve norms and certain other geometric properties upon dimensionality reduction.[2] In particular, our testers are *oblivious* to the input. It is a very interesting open question as to whether non-oblivious testers can strengthen the above results.

1.1. Our Results

We now present our results more formally. For integer $m > 0$, let $\mathcal{S}^{m-1} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\| = 1\}$, and let $\mathsf{Sp}_k^m = \{\mathbf{x} \in \mathcal{S}^{m-1} : \|\mathbf{x}\|_0 \leqslant k\}$.[3]

Theorem 1.2 (Unknown Design Matrix). *Fix $\varepsilon, \delta \in (0,1)$ and positive integers d, k, m and p, such that $(k/m)^{1/8} < \varepsilon < \frac{1}{100}$ and $k \geqslant 10\log\frac{1}{\varepsilon}$. There exists a tester with query complexity $O(\varepsilon^{-2}\log(p/\delta))$ which, given as input vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p \in \mathbb{R}^d$, has the following behavior (where $\mathbf{Y}$ is the matrix having $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$ as columns):*

- **Completeness**: *If $\mathbf{Y}$ admits a decomposition $\mathbf{Y} = \mathbf{A}\mathbf{X}$, where $\mathbf{A} \in \mathbb{R}^{d\times m}$ satisfies (ε, k)-RIP and $\mathbf{X} \in \mathbb{R}^{m\times p}$ with each column of $\mathbf{X}$ in Sp_k^m, then the tester accepts with probability $\geqslant 1-\delta$.*

- **Soundness**: *Suppose $\mathbf{Y}$ does not admit a decomposition $\mathbf{Y} = \mathbf{A}(\mathbf{X}+\mathbf{Z}) + \mathbf{W}$ with*

  1. *The design matrix $\mathbf{A} \in \mathbb{R}^{d\times m}$ being (ε, k)-RIP, with $\|\mathbf{a}_i\| = 1$ for every $i \in [m]$.*
  2. *The coefficient matrix $\mathbf{X} \in \mathbb{R}^{m\times p}$ being column wise ℓ-sparse, where $\ell = O(k/\varepsilon^4)$.*
  3. *The error matrices $\mathbf{Z} \in \mathbb{R}^{m\times p}$ and $\mathbf{W} \in \mathbb{R}^{d\times p}$ satisfying*

$$\|\mathbf{z}_i\|_\infty \leqslant \varepsilon^2, \qquad \|\mathbf{w}_i\|_2 \leqslant O(\varepsilon^{1/4}) \qquad \textit{for all } i \in [p].$$

  Then the tester rejects with probability $\geqslant 1-\delta$.

[1] With slight abuse of notation, we use the term basis to denote the set of columns of a design matrix. The columns might not be linearly independent.

[2] If evaluating the queries efficiently was an objective, one could also use sparse dimension reduction matrices (Dasgupta et al., 2010; Kane & Nelson, 2014; Bourgain et al., 2015), but we do not pursue this direction here.

[3] Here, $\|\mathbf{x}\|_0$ denotes the the sparsity of the vector, $\|\mathbf{x}\|_0 := |\{i \in [m] \mid x_i \neq 0\}|$. Without any subscript, $\|\cdot\|$ denotes the ℓ_2-norm: $\|\mathbf{x}\| := \sqrt{\sum_i x_i^2}$.

The contrapositive of the soundness guarantee from the above theorem states that if the tester accepts, then matrix $\mathbf{Y}$ admits a factorization of the form $\mathbf{Y} = \mathbf{A}(\mathbf{X}+\mathbf{Z})+\mathbf{W}$, with error matrices $\mathbf{Z}$ and $\mathbf{W}$ having ℓ_∞ and ℓ_2 error bounds. The matrix $\mathbf{X}+\mathbf{Z}$ is a sparse matrix with ℓ_∞-based thresholding, and $\mathbf{W}$ is an additive ℓ_2-error term.[4]

Theorem 1.3 (Known Design Matrix)**.** *Fix $\varepsilon, \delta \in (0,1)$ and positive integers d, k, m and a matrix $\mathbf{A} \in \mathbb{R}^{d\times m}$ such that $\|\mathbf{a}_i\| = 1$ for every $i \in [m]$. There exists a tester with query complexity $O(k\varepsilon^{-2}\log(m/\delta))$ that behaves as follows for an input vector $\mathbf{y} \in \mathbb{R}^d$:*

- **Completeness***: If $\mathbf{y} = \mathbf{A}\mathbf{x}$ for some $\mathbf{x} \in \mathsf{Sp}_k^m$, then the tester accepts with probability* 1.
- **Soundness***: If $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 > \varepsilon$ for every $\mathbf{x} : \|\mathbf{x}\|_0 \leqslant K$, then the tester rejects with probability $\geqslant 1-\delta$. Here, $K = O(k/\varepsilon^2)$.*

The running time of the tester is $\mathrm{poly}(m, k, 1/\varepsilon)$.

A different way of stating the result is that the tester, using $O(k\varepsilon^{-2}\log(m/\delta))$ linear queries, accepts with probability 1 if $\mathbf{y} = \mathbf{A}\mathbf{x}$ for a k-sparse $\mathbf{x} \in \mathbb{R}^m$ and rejects with probability $1-\delta$ if $\|\mathbf{A}\mathbf{x} - \mathbf{y}\| > \varepsilon\|\mathbf{x}\|$ for every $O(k/\varepsilon^2)$-sparse $\mathbf{x}$. To complement this result, we show that a better tradeoff between the sparsity and reconstruction error is likely to be impossible.

Theorem 1.4 (Hardness)**.** *Assume* SAT *does not have $n^{O(\log\log n)}$-time algorithms, and let η be any constant less than* 1*. Then, there does not exist a polynomial time algorithm that, given input $\mathbf{A} \in \mathbb{R}^{d\times m}$ (where $\|\mathbf{a}_i\| = 1$ for every $i \in [m]$), $\mathbf{y} \in \mathbb{R}^d$ and $\varepsilon > 0$, distinguishes with constant probability between the following two cases: (i) $\mathbf{y} = \mathbf{A}\mathbf{x}$ for a k-sparse $\mathbf{x}$, and (ii) $\|\mathbf{y} - \mathbf{A}\mathbf{x}\| > \varepsilon\|\mathbf{x}\|^\eta$ for every (k/ε^2)-sparse $\mathbf{x}$.*

Note that the above hardness applies to any polynomial time algorithm, not just sketching algorithms.

We also give *tolerant* variants of these testers (Theorems H.1 and H.2) which can handle bounded noise for the completeness case. Moreover, the tester for the known design case can be converted into a new sketching algorithm for *sparse recovery* (Theorem D.1).

Finally, we also give an algorithm for testing dimensionality, which is based on similar techniques.

Theorem 1.5 (Testing Dimensionality)**.** *Fix $\varepsilon, \delta \in (0,1)$, positive integers d, k and p, where $k \geqslant 10\varepsilon^2\log d$. There exists a tester with query complexity $O(p\log\delta^{-1})$, which gives as input vectors $\mathbf{y}_1, \ldots, \mathbf{y}_p \subset \mathcal{S}^{d-1}$, has the following behavior:*

- **Completeness***: If $\mathrm{rank}(Y) \leqslant k$, then the tester accepts with probability $\geqslant 1-\delta$.*
- **Soundness***: If $\mathrm{rank}_\varepsilon(Y) \geqslant k'$, then the tester rejects with probability $\geqslant 1-\delta$. Here, $k' = 20k/\varepsilon^2$*

The soundness criteria in the above Theorem is stated in terms of the ε-approximate rank of a matrix (see Definition E.1). This is a well-studied relaxation of the standard definition of rank, and has applications in approximation algorithms, communication complexity and learning theory (see (Alon et al., 2013) and references therein).

1.2. Related Work

Although, to the best of our knowledge, the testing problems we consider have not been explicitly investigated before, there are several related areas of study that frame our results in their proper context.

Unknown Design setting. In the setting of the unknown design matrix, the question of recovering the design matrix and the sparse representation (as opposed to our problem of testing their existence) is called the *dictionary learning* or *sparse coding* problem. The first work to give a dictionary learning algorithm with provable guarantees was (Spielman et al., 2012) where the dictionary was restricted to be square. For the more common overcomplete setting, (Arora et al., 2014) and (Agarwal et al., 2014) independently gave algorithms with provable guarantees for dictionaries satisfying incoherence and RIP respectively. All of these (as well as other more recent) works assume distributions from which the input samples are generated in an i.i.d fashion. In contrast, our work is in the *agnostic setting* and hence, is incomparable with these results.

It is known that the dictionary learning problem is NP-hard, even for square dictionaries (Razaviyayn et al., 2014; Tillmann, 2015). In fact, (Tillmann, 2015) shows that unless SAT has a quasi-polynomial time algorithm, it is impossible, given $\mathbf{Y} \in \mathbb{R}^{d\times p}$, to approximate in polynomial time the minimum k upto a factor $2^{\log^{1-\varepsilon} d}$ (for any $\varepsilon > 0$) such that $\mathbf{Y} = \mathbf{A}\mathbf{X}$ where each column of $\mathbf{X} \in \mathbb{R}^{d\times p}$ is k-sparse. This motivates our bicriteria relaxation of both the sparsity as well as the additive error in Theorem 1.2.

Known Design setting. Some results about testing sparsity in the known design setting are implicit in recent work on streaming algorithms and oblivious subspace embeddings. Of particular interest are the following results:

Theorem 1.6 (Implicit in (Kane et al., 2010))**.** *Fix $\varepsilon \in (0,1)$, positive integers m, k and an* invertible *matrix* $\mathbf{A} \in$

[4] Theorem 1.2 can be restated in terms of *incoherent* (instead of RIP) design matrices as well. This follows from the fact that the incoherence and RIP constants of a matrix are order-wise equivalent. This observation is formalized in Appendix F.

$\mathbb{R}^{m\times m}$. Then, there is a tester with query complexity $O(\varepsilon^{-2}\log(m))$ that, for an input $\mathbf{y}\in\mathbb{R}^m$, accepts with probability at least $2/3$ if $\mathbf{y}=\mathbf{A}\mathbf{x}$ for some k-sparse $\mathbf{x}\in\mathbb{Z}^m$, and rejects with probability $2/3$ if $\mathbf{y}\neq\mathbf{A}\mathbf{x}$ for all $(1+\varepsilon)k$-sparse $\mathbf{x}\in\mathbb{Z}^m$. The running time of the algorithm is $\operatorname{poly}(m,1/\varepsilon)$.

Theorem 1.7 (Implicit in prior work, see (Woodruff, 2014))**.** *Fix $\varepsilon,\delta\in(0,1)$ and positive integers d,k,m and a matrix $\mathbf{A}\in\mathbb{R}^{d\times m}$. Then, there is a tester with query complexity $O(k\varepsilon^{-2}\log(m/\delta))$ that, for an input vector $\mathbf{y}\in\mathbb{R}^d$, accepts with probability 1 if $\mathbf{y}=\mathbf{A}\mathbf{x}$ for some k-sparse $\mathbf{x}$ and rejects with probability at least $1-\delta$ if $\|\mathbf{y}-\mathbf{A}\mathbf{x}\|>\varepsilon$ for all k-sparse $\mathbf{x}$. The running time of the tester is the time required to solve the following optimization problem:*

$$\widehat{\mathbf{x}}=\arg\min_{\mathbf{x}'\in K}\|\mathbf{S}\mathbf{A}\mathbf{x}'-\mathbf{S}\mathbf{y}\|=\arg\min_{\mathbf{x}'\in K}\|\mathbf{S}(\mathbf{A}\mathbf{x}'-\mathbf{y})\| \tag{1}$$

where $\mathbf{S}\in\mathbb{R}^{q\times d}$ is a random sketch matrix (where $q\ll d$) and $K=\{\mathbf{x}:\|\mathbf{x}\|_0\leqslant k\}$

Detailed descriptions of the algorithms and proof sketches for the above Theorems are given in Section B.4. The algorithms from the above theorems come with significant limitations. In particular, the guarantees for Theorem 1.6 hold only when the design matrix is invertible. On the other hand, the running time for the algorithm in Theorem 1.7 is the cost of solving the optimization problem in Equation (1), which is known to be NP-hard for general matrices.

The problem of testing sparsity has also been studied in *non-sketching* settings as well, where the algorithm is allowed access to the entire input. In particular, (Natarajan, 1995) gave a bicriteria-approximation algorithm, where the blowup in the sparsity is proportional to $\|\mathbf{A}^\dagger\|_2^2$ (which can be large if $\mathbf{A}$ is ill conditioned).

Testing Dimensionality. In (Czumaj et al., 2000), some problems in computational geometry were studied from the property testing perspective, but the problems involved only discrete structures. (Krauthgamer & Sasson, 2003) studied the problem of testing dimensionality, but their notion of farness from being low-dimensional is different from ours[5]. (Chierichetti et al., 2017) gave approximation algorithms for computing approximate rank of the matrix, in the setting where the algorithms have *full access* to the input.

1.3. Discussion

A standard approach to designing a testing algorithm for a property $\mathcal{P}$ is the following: we identify an alternative property $\mathcal{P}'$ which can be *tested efficiently and exactly*, while satisfying the following:

(i) **Completeness**: If an instance satisfies $\mathcal{P}$, then it satisfies $\mathcal{P}'$.

(ii) **Soundness**: If an instance satisfies $\mathcal{P}'$, the it is close to satisfying $\mathcal{P}$.

In other words, we reduce the property testing problem to that of finding a efficiently testable property $\mathcal{P}'$, which can be interpreted as a surrogate for property $\mathcal{P}$. The inherent geometric nature of the problems looked at in this paper motivate us to look for $\mathcal{P}'$s which are based around convex geometry and high dimensional probability.

For the unknown design setting, we are intuitively looking for a $\mathcal{P}'$ based on a quantity ω that *robustly* captures sparsity and is easily computable using linear queries, in the sense that ω is small when the input vectors have a sparse coding and large when they are "far" from any sparse coding. Moreover, ω needs to be invariant with respect to isometries and nearly invariant with respect to near-isometries. A natural and widely-used measure of structure that satisfies the above mentioned properties is the *gaussian width.*

Definition 1.8. *The* gaussian width *of a set $S\subseteq\mathbb{R}^d$ is: $\omega(S)=\mathbf{E}_{\mathbf{g}}[\sup_{\mathbf{v}\in S}\langle\mathbf{g},\mathbf{v}\rangle]$ where $\mathbf{g}\in\mathbb{R}^d$ is a random vector drawn from $N(0,1)^d$, i.e., a vector of independent standard normal variables.*

The gaussian width of S measures how well on average the vectors in S correlate with a randomly chosen direction. It is invariant under orthogonal transformations of S as the distribution of $\mathbf{g}$ is spherically symmetric. It is a well-studied quantity in high-dimensional geometry ((Vershynin, 2015; Mendelson & Vershynin, 2002)), optimization ((Chandrasekaran et al., 2012; Amelunxen et al., 2013)) and statistical learning theory ((Bartlett & Mendelson, 2002)). The following bounds are well-known.

Lemma 1.9 (See, for example, (Rudelson & Vershynin, 2008; Vershynin, 2015))**.**

(i) If S is a finite subset of $\mathcal{S}^{d-1}$, then $\omega(S)\leqslant\sqrt{2\log|S|}$.

(ii) $\omega(\mathcal{S}^{d-1})\leqslant\sqrt{d}$

(iii) If $S\subseteq\mathcal{S}^{d-1}$ is of dimension k, then $\omega(S)\leqslant\sqrt{k}$.

(iv) $\omega(\mathsf{Sp}_k^d)\leqslant 2\sqrt{3k\log(d/k)}$ when $d/k>2$ and $k\geqslant 4$.

In the context of Theorems 1.2 and 1.5, one can observe that whenever a given set satisfies sparsity or dimensionality constraints, the gaussian width of such sets are small (points (iii) and (iv) from the above Lemma). Therefore, one can hope to test dimensionality or sparsity by computing an empirical estimate of the gaussian width and comparing the estimate to the results in Lemma 1.9. While completeness of such testers would follow directly from concentration of measure, establishing soundness would require us to show

[5] In their setup, a sequence of vectors $\mathbf{y}_1,\dots,\mathbf{y}_p$ is ε-far from being d-dimensional if at least εp vectors need to be removed to make it be of dimension d

that approximate converses of points (iii) and (iv) hold as well i.e., whenever the gaussian width of the set S is small, it can be approximated by sets which are approximately sparse in some design matrix (or have low rank).

For the soundness direction of Theorem 1.2, the above arguments are made precise using Lemma 3.3 and Theorem 3.2, which show that small gaussian width sets can be approximated by random projections of sparse vectors and vectors with small ℓ_∞-norm. For Theorem 1.5, we use lemma E.2 which shows that sets with small gaussian width have small approximate rank.

For the known design setting, we are looking for a $\mathcal{P}'$, which would ensure that if a given point $\mathbf{y} \in \mathbb{R}^d$ satisfies $\mathcal{P}'$, then it is close to having a sparse representation in the matrix $\mathbf{A}$. Towards this end, the approximate Carathéodory's theorem states that if a point $\mathbf{y} \in \mathbb{R}^d$ belonging to the convex-hull of $\mathbf{A}$, then it is close to another point which admits a sparse representation. On the other hand, if a unit vector $\mathbf{x} \in \mathcal{S}^{d-1} \cap \mathbb{R}^d_+$ were k-sparse to begin with , then it can be seen that the corresponding $\mathbf{y} = \mathbf{A}\mathbf{x}$ would belong to the convex hull of $\sqrt{k} \cdot \mathbf{A}$. These observations taken together, seem to suggest that one can take $\mathcal{P}'$ to be membership in the convex-hull of $\sqrt{k} \cdot \mathbf{A}$. This intuition is made precise in the analysis of the tester in Section 4.

1.4. Organization

Section 2 introduces notations and preliminaries used in the rest of the paper. In Sections 3 and 4, we design and analyze the testers for the unknown and known basis setting respectively. Section 5 contains empirical results which supplement Section 3. In Section B we prove additional lemmas used in the proof of Theorem 3.2, and in Section A we prove Theorem 3.2. In Section C, we prove Theorem C.1, a stronger version of Theorem 1.4. In Section D, we show that Theorem 1.3 yields a sketching algorithm for sparse recovery. In Section E, we design and analyze the dimensionality tester. In Section G, we describe the results for testing sparsity in the known case implicit in previous work. Finally, in Section H, we give noise tolerant testers for the known and unknown basis settings.

2. Preliminaries

Given $S \subset \mathbb{R}^d$, we shall use $\mathrm{conv}(S)$ to denote the convex hull of S. For a vector $\mathbf{x} \in \mathbb{R}^d$, we use $\|\cdot\|_p$ to denote its ℓ_p-norm, and we will drop the indexing when $p = 2$. We denote the ℓ_2-distance of the point $\mathbf{x}$ to the set S by $\mathrm{dist}(\mathbf{x}, S)$. We recall the definition of ε-isometry:

Definition 2.1. *Given sets $S \subset \mathbb{R}^m$ and $S' \subset \mathbb{R}^n$ (for some $m, n \in \mathbb{N}$), we say that S' is an ε-isometry of S, if there exists a mapping $\psi : S \mapsto S'$ which satisfies the following property:*

$$\forall \mathbf{x}, \mathbf{y} \in S : (1-\varepsilon)\|\mathbf{x}-\mathbf{y}\| \leqslant \|\psi(\mathbf{x})-\psi(\mathbf{y})\| \leqslant (1+\varepsilon)\|\mathbf{x}-\mathbf{y}\|$$

For the unknown design setting, we shall require the notion of Restricted Isometry Property, which is defined as follows:

Definition 2.2 ((ε, k)-RIP)**.** *A matrix $\mathbf{A} \in \mathbb{R}^{d \times m}$ satisfies (ε, k)-RIP, if for every $\mathbf{x} \in \mathsf{Sp}^m_k$ the following holds:*

$$(1-\varepsilon)\|\mathbf{x}\| \leqslant \|\mathbf{A}\mathbf{x}\| \leqslant (1+\varepsilon)\|\mathbf{x}\| \tag{2}$$

We use the following version of Gordon's Theorem repeatedly in this work.

Theorem 2.3 (Gordon's Theorem (Gordon, 1985))**.** *Given $S \subset \mathcal{S}^{D-1}$ and a random gaussian matrix $\mathbf{G} \sim \frac{1}{\sqrt{d'}} N(0,1)^{d' \times D}$, we have*

$$\mathop{\mathbf{E}}_{\mathbf{G}}\left[\max_{\mathbf{x} \in S} \|\mathbf{G}\mathbf{x}\|_2\right] \leqslant 1 + \frac{\omega(S)}{\sqrt{d'}}$$

It directly implies the following generalization of the Johnson-Lindenstrauss lemma.

Theorem 2.4 (Generalized Johnson-Lindenstrauss lemma)**.** *Let $S \subseteq \mathcal{S}^{n-1}$. Then there exists linear transformation $\Phi : \mathbb{R}^n \mapsto \mathbb{R}^{d'}$, for $d' = O\left(\frac{\omega(S)^2}{\varepsilon^2}\right)$, such that Φ is an ε-isometry on S. Moreover, $\Phi \sim \frac{1}{\sqrt{d'}} N(0,1)^{d' \times n}$ is an ε-isometry on S with high probability.*

It can be easily verified that the quantity $\max_{\mathbf{x} \in S} \|\mathbf{G}\mathbf{x}\|_2$ is 1-Lipschitz with respect to $\mathbf{G}$. Therefore, using Gaussian concentration for Lipschitz functions, we get the following corollary :

Corollary 2.5. *Let S and G be as in Theorem 2.3. Then for all $\varepsilon > 0$, we have*

$$\begin{aligned} \mathop{\mathbf{Pr}}_{\mathbf{G}}\left(\max_{\mathbf{x} \in S} \|\mathbf{G}\mathbf{x}\|_2 \geqslant 1 \right. & + \left. (1+\varepsilon)\frac{\omega(S)}{\sqrt{d'}}\right) \\ & \leqslant \exp\left(-O(\varepsilon\omega(S))^2\right) \end{aligned}$$

The following lemma gives concentration for the gaussian width:

Lemma 2.6 (Concentration on the gaussian width (Boucheron et al., 2013))**.** *Let $S \subset \mathbb{R}^d$. Let $W = \sup_{\mathbf{v} \in S}\langle \mathbf{g}, \mathbf{v}\rangle$ where $\mathbf{g}$ is drawn from $N(0,1)^d$. Then:*

$$\mathbf{Pr}[|W - \mathbf{E}\, W| > u] < 2e^{-\frac{u^2}{2\sigma^2}}$$

where $\sigma^2 = \sup_{\mathbf{v} \in S}\left(\|\mathbf{v}\|_2^2\right)$. Notice that the bound is dimension independent.

Lastly, we shall use the ℓ_2-variant of the approximate Carathéodory's Theorem:

Theorem 2.7. *(Theorem* 0.1.2 *(Vershynin, 2016)) Given $X = \{\mathbf{w}_1, \dots, \mathbf{w}_p\}$ where $\|\mathbf{w}_i\| \leqslant 1$ for every $i \in [p]$. Then for every choice $\mathbf{z} \in \operatorname{conv}(X)$ and $k \in \mathbb{N}$, there exists $\mathbf{w}_{i_1}, \mathbf{w}_{i_2}, \dots, \mathbf{w}_{i_k}$ such that*

$$\left\| \frac{1}{k} \sum_{j \in [k]} \mathbf{w}_{i_j} - \mathbf{z} \right\| \leqslant \frac{2}{\sqrt{k}} \tag{3}$$

2.1. Algorithmic Estimation of Gaussian Width and Norm of a vector

We record here simple lemmas bounding the number of linear queries needed to estimate the gaussian width of a set and the length of a vector.

Lemma 2.8 (Estimating Gaussian Width using linear queries)**.** *For any $u > 4$, $\varepsilon \in (0, 1/2)$ and $\delta > 0$, there is a randomized algorithm that given a set $S \subseteq \mathbb{R}^d$ and $\|\mathbf{v}\| \in [1 \pm \varepsilon]$ for all $\mathbf{v} \in S$, computes $\hat{\omega}$ such that $\omega(S) - u \leqslant \hat{\omega} \leqslant \omega(S) + u$ with probability at least $1 - \delta$. The algorithm makes $O(\log(1/\delta) \cdot |S|)$ linear queries to S.*

Proof. By Lemma 2.6, for a random $\mathbf{g} \sim N(0, 1)^d$, $\sup_{\mathbf{v} \in S} \langle \mathbf{g}, \mathbf{v} \rangle$ is away from $\omega(S)$ by u with probability at most $2e^{-16/4.5} < 0.1$. By the Chernoff bound, the median of $O(\log \delta^{-1})$ trials will satisfy the conditions required of $\hat{\omega}$ with probability at least $1 - \delta$. □

Lemma 2.9 (Estimating norm using linear queries)**.** *Given $\varepsilon \in (0, 1/2)$ and $\delta > 0$, for any vector $\mathbf{x} \in \mathbb{R}^d$, only $O(\varepsilon^{-2} \log \delta^{-1})$ linear queries to $\mathbf{x}$ suffice to decide whether $\|\mathbf{x}\| \in [1 - \varepsilon, 1 + \varepsilon]$ with success probability $1 - \delta$.*

Proof. It is easy to verify that $\mathbf{E}_{\mathbf{g} \sim N(0,1)^d}[\langle \mathbf{g}, \mathbf{x} \rangle^2] = \|\mathbf{x}\|^2$. Therefore, it can be estimated to a multiplicative error of $(1 \pm \varepsilon/2)$ by taking the average of the squares of linear measurements using $O\left(\frac{1}{\varepsilon^2} \log \frac{1}{\delta}\right)$-queries. For the case $\|\mathbf{x}\|_2 \leqslant 2$, a multiplicative error $(1 \pm \varepsilon/2)$ implies an additive error of ε. Furthermore, when $\|\mathbf{x}\|_2 \geqslant 2$, a multiplicative error of $(1 \pm \varepsilon/2)$ implies that $L \geqslant 2(1 - \varepsilon/2) > 1 + \varepsilon$ for $\varepsilon < 1/2$. □

3. Analysis for Unknown Design setting

In this section, we prove Theorem 1.2. Let S denote the set $\{\mathbf{y}_1, \dots, \mathbf{y}_p\}$. Our testing algorithm is shown in Algorithm 1.

The number of linear queries made by the tester is $O(p\varepsilon^{-2} \log(p/\delta))$ in Line 1 and $O(p \log \delta^{-1})$ in Line 2.

3.1. Completeness

Assume that for each $i \in [p]$, $\mathbf{y}_i = A\mathbf{x}_i$ for a matrix $A \in \mathbb{R}^{d \times m}$ satisfying (ε, k)-RIP and $\mathbf{x}_i \in \mathsf{Sp}_k^m$. By definition of RIP, we know that $1 - \varepsilon \leqslant \|\mathbf{y}_i\| \leqslant 1 + \varepsilon$, so that Line 1 of the algorithm will pass with probability at least $1 - \delta/2$.

Algorithm 1 SparseTestUnknown

```
1: Use Lemma 2.9 to decide with probability at least $1 - \delta/2$ if there exists $\mathbf{y}_i$ such that $\|\mathbf{y}_i\| \notin [1 - 2\varepsilon, 1 + 2\varepsilon]$. Reject if so.
2: Use Lemma 2.8 to obtain $\hat{\omega}$, an estimate of $\omega(S)$ within additive error $\sqrt{3k \log(m/k)}$ with probability at least $1 - \delta/2$.
3: Accept if $\hat{\omega} \leqslant 4\sqrt{3k \log(m/k)}$, else reject.
```

From Lemma 1.9, we know that $\omega(\{\mathbf{x}_1, \dots \mathbf{x}_p\}) \leqslant 2\sqrt{3k \log(m/k)}$. Lemma 3.1 shows that the gaussian width of S is approximately the same; its proof, deferred to the appendix (Section B.4), uses Slepian's Lemma (Lemma B.3).

Lemma 3.1. *Let $X \subset \mathcal{S}^{m-1}$ be a finite set, and let $S \subset \mathbb{R}^d$ be an ε-isometric embedding of X. Then*

$$(1 - \varepsilon)\omega(X) \leqslant \omega(S) \leqslant (1 + \varepsilon)\omega(X) \tag{4}$$

Hence, the gaussian width of $\mathbf{y}_1, \dots, \mathbf{y}_p$ is at most $2(1 + \varepsilon)\sqrt{3k \log(m/k)}$. Taking into account the additive error in Line 2, we see that with probability at least $1 - \delta/2$, $\hat{\omega} \leqslant (3 + 2\varepsilon)\sqrt{3k \log(m/k)} \leqslant 4\sqrt{3k \log(m/k)}$. Hence, the tester accepts with probability at least $1 - \delta$.

3.2. Soundness

As mentioned before, in order to prove soundness we need to show that whenever the gaussian width of the set S is small, it is *close* to some sparse point-set. Let $\omega^* = 4\sqrt{3k \log \frac{m}{k}}$. We shall break the analysis into two cases:

Case (i) $\left\{\omega^* \geqslant (\varepsilon/C)^2 \sqrt{d}\right\}$: For this case, we use the fact random projection of discretized sparse point-sets (Definition A.1) form an appropriated cover of S. This is formalized in the following theorem, which in a sense shows an approximate inverse of Gordon's Theorem for sparse vectors:

Theorem 3.2. *Given $\varepsilon > 0$ and integers C, d, k and m, let $n = O\left(\frac{k}{\varepsilon^2} \log(m/k)\right)$. Suppose $m \geqslant k/\varepsilon^8$. Let $\Phi : \mathbb{R}^m \mapsto \mathbb{R}^n$ be drawn from $\frac{1}{\sqrt{n}} N(0, 1)^{n \times m}$. Then, for $\ell = O(k\varepsilon^{-4})$, with high probability, the set $\Phi^{\mathrm{norm}}(\widehat{\mathsf{Sp}_\ell^m})$ is an $O(\varepsilon^{1/4})$-cover of $\mathcal{S}^{n-1}$, where $\Phi^{\mathrm{norm}}(\mathbf{x}) = \Phi(\mathbf{x})/\|\Phi(\mathbf{x})\|_2$.*

The proof of the above Theorem is deferred to Section A. From the choice of parameters we have $d \leqslant \frac{C'k}{\varepsilon^2} \log \frac{m}{k}$ Therefore, using the above Theorem we know that there exists (ε, k)-RIP matrix $\Phi \in \mathbb{R}^{d\times m}$ such that $\Phi^{\text{norm}}\big(\mathsf{Sp}_\ell^m\big)$ is an $O(\varepsilon^{1/4})$-cover of $\mathcal{S}^{d-1}$ (and therefore it is a $\varepsilon^{1/4}$-cover of S). Therefore, there exists $\mathbf{X} \in \mathbb{R}^{m\times p}$ such that $\mathbf{Y} = \Phi(\mathbf{X}) + \mathbf{E}$ where the columns of $\mathbf{X}$ and $\mathbf{E}$ satisfy the respective $\|\cdot\|_0$ and $\|\cdot\|_2$-upper bounds respectively.

Case (ii) $\left\{\omega^* \leqslant (\varepsilon/C)^2\sqrt{d}\right\}$: For this case, we use the following result on the concentration of ℓ_∞-norm:

Lemma 3.3. *Given $S \subset \mathcal{S}^{d-1}$, we have*

$$\Pr_{\mathbf{R}\sim\mathbb{O}_d}\left[\max_{\mathbf{y}\in\mathbf{R}(S)} \|\mathbf{y}\|_\infty \leqslant C\frac{\omega(S)}{d^{1/2}}\right] \geqslant \frac{1}{2}$$

where $\mathbb{O}_d$ is the orthogonal group in $\mathbb{R}^d$ i.e., $\mathbf{R}$ is a uniform random rotation.

Although this concentration bound is known, for completeness we give a proof in the appendix (Section B.7). From the above lemma, it follows that there exists $\mathbf{R} \in \mathbb{O}_d$ such that for any $\mathbf{z} \in Z := \mathbf{R}(S)$ we have $\|\mathbf{z}\|_\infty \leqslant \varepsilon^2$ and therefore $\mathbf{Y} = \mathbf{R}^{-1}\mathbf{Z}$. Furthermore, since $\mathbf{R}$ is orthogonal, therefore the matrix $\mathbf{R}^{-1}$ is also orthogonal, and therefore it satisfies (ε, k)-RIP.

To complete the proof, we observe that even though the given factorization has inner dimension d, we can trivially extend it to one with inner dimension m. This can be done by constructing $\Phi = \begin{bmatrix}\mathbf{R}^{-1} & \mathbf{G}\end{bmatrix}$ with $\mathbf{G} \sim \frac{1}{\sqrt{d}}N(0,1)^{d\times m-d}$. Since $\omega^* \lll d$, from Theorem 2.4 it follows that with high probability $\mathbf{G}$ (and consequently Φ) will satisfy (ε, k)-RIP. Finally, we construct $\hat{\mathbf{Z}} \in \mathbb{R}^{m\times n}$ by padding $\mathbf{Z}$ with $m-d$ rows of zeros. Therefore, by construction $Y = \Phi \cdot \hat{\mathbf{Z}}$, where for every $i \in [p]$ we have $\|\mathbf{z}_i\|_\infty \leqslant \varepsilon^2$. Hence the claim follows.

4. Analysis for the Known Design setting

In this section, we describe and analyze the tester for the known design matrix case. The algorithm itself is a simple convex-hull membership test, which can be solved using a linear program.

Algorithm 2 SparseTest-KnownDesign

```
Set $n = 100k\log\frac{m}{\delta}$, sample projection matrix $\Phi \sim \frac{1}{\sqrt{n}}N(0,1)^{n\times d}$
Observe linear sketch $\tilde{\mathbf{y}} = \Phi(\mathbf{y})$
Let $A_\pm = A \cup -A$
Accept iff $\tilde{\mathbf{y}} \in \sqrt{k}\cdot\mathrm{conv}\big(\Phi(A_\pm)\big)$
```

We shall now prove the completeness and soundness guarantees of the above tester. The running time bound follows because convex hull membership reduces to linear programming.

Completeness Let $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{d\times m}$ is an arbitrary matrix with $\|\mathbf{a}_i\| = 1$ for every $i \in [m]$. Furthermore $\|\mathbf{x}\|_2 = 1$ and $\|\mathbf{x}\|_0 \leqslant k$. Therefore, by Cauchy-Schwartz we have $\|\mathbf{x}\|_1 \leqslant \sqrt{k}\|\mathbf{x}\|_2 = \sqrt{k}$. Hence, it follows that $\mathbf{y} \in \sqrt{k}\cdot\mathrm{conv}(A_\pm)$. Since $\Phi : \mathbb{R}^m \mapsto \mathbb{R}^d$ is a linear transformation, we have $\Phi(\mathbf{y}) \in \sqrt{k}\cdot\mathrm{conv}(\Phi(A_\pm))$. Therefore, the tester accepts with probability 1.

Soundness Consider the set $A_{\varepsilon/\sqrt{k}}$ which is the set of all $(2k/\varepsilon^2)$-uniform convex combinations of $\sqrt{k}(A_\pm)$ i.e.,

$$A_{\varepsilon/\sqrt{k}} = \left\{\sum_{\mathbf{v}_i\in\Omega}\frac{\varepsilon^2}{2k}\mathbf{v}_i : \text{multiset } \Omega \in \Big(\sqrt{k}.A_\pm\Big)^{2k/\varepsilon^2}\right\} \tag{5}$$

Then, from the approximate Carathéodory theorem, it follows that $A_{\varepsilon/\sqrt{k}}$ is an ε-cover of $\sqrt{k}\cdot\mathrm{conv}\big(A_\pm\big)$. Furthermore, $|A_{\varepsilon/\sqrt{k}}| \leqslant (2m)^{2k/\varepsilon^2}$. By our choice of n, with probability at least $1-\delta/2$, the set $\Phi\Big(\{\mathbf{y}\}\cup A_{\varepsilon/\sqrt{k}}\Big)$ is ε-isometric to $\{\mathbf{y}\}\cup A_{\varepsilon/\sqrt{k}}$.

Let $\tilde{A}_{\varepsilon/\sqrt{k}} = \Phi\big(A_{\varepsilon/\sqrt{k}}\big)$. Again, by the approximate Carathéodory's theorem, the set $\tilde{A}_{\varepsilon/\sqrt{k}}$ is an ε-cover of $\Phi\big(\sqrt{k}\cdot\mathrm{conv}(A_\pm)\big)$. Now suppose the test accepts $\mathbf{y}$ with probability at least δ. Then, with probability at least $\delta/2$, the test accepts and the above ε-isometry conditions hold simultaneously. Then,

$$\begin{aligned}
&\tilde{\mathbf{y}} \in \sqrt{k}\cdot\mathrm{conv}\big(\Phi(A_\pm)\big)\\
\overset{1}{\Rightarrow}\ &\mathrm{dist}\big(\tilde{\mathbf{y}}, \tilde{A}_{\varepsilon/\sqrt{k}}\big) \leqslant \varepsilon\\
\overset{2}{\Rightarrow}\ &\mathrm{dist}\big(\mathbf{y}, A_{\varepsilon/\sqrt{k}}\big) \leqslant \varepsilon(1-\varepsilon)^{-1} \leqslant 2\varepsilon\\
\Rightarrow\ &\mathrm{dist}\big(\mathbf{y}, \sqrt{k}\cdot\mathrm{conv}(A_\pm)\big) \leqslant 2\varepsilon
\end{aligned}$$

where step 1 follows from the ε-cover guarantee of $\tilde{A}_{\varepsilon/\sqrt{k}}$, step 2 follows from the ε-isometry guarantee. Invoking the approximate Carathéodory theorem, we get that there exists $\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}} \in \sqrt{k}\cdot\mathrm{conv}(\pm A)$ such that $\|\hat{\mathbf{x}}\|_0 \leqslant O(k/\varepsilon^2)$ and $\|\hat{\mathbf{y}} - \mathbf{y}\| \leqslant O(\varepsilon)$. This completes the soundness direction.

5. Experimental Results

Our algorithm for the unknown design setting is based on the principle that the property of sparse representability in some basis admits an approximate characterization in terms of gaussian width. This section provides experimental evidence which supplements our theoretical results. For the empirical study, we use the classic *Barbara* image (which is of size 512×512 pixels). Specifically, we consider 9 sub-images of size 100×100 pixels each (see Figure 1). For each such sub-image, we compute a matrix representation (by the standard technique of subdividing the images into patches, see, e.g., (Elad & Aharon, 2006)). In particular,

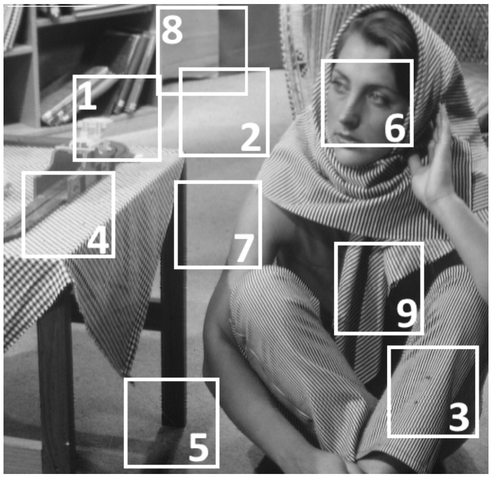

Figure 1. Sub-images used as data points.

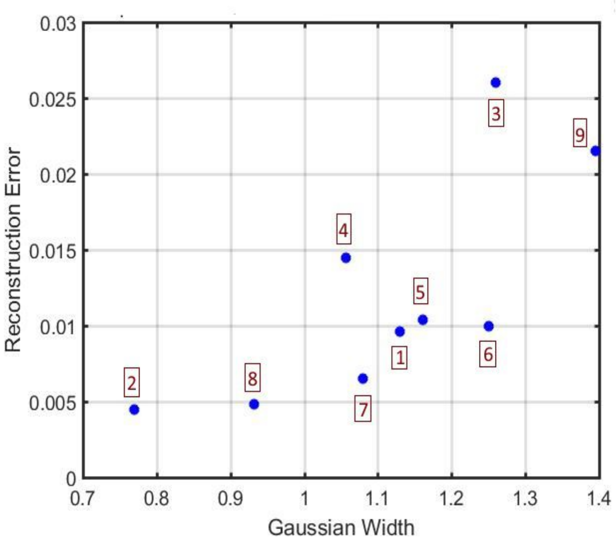

Figure 2. Correlation of gaussian width and reconstruction error.

each sub-image is represented as a matrix $\mathbf{Y}$ of dimension 64×8649. Then, for each matrix $\mathbf{Y}$ corresponding to a sub-image, we estimate the gaussian width of the ℓ_2-column normalized matrix. In addition, setting the number of atoms $m = 100$ and sparsity $k = 10$, we run the k-SVD algorithm for 50 iterations and record the reconstruction error.[6]

Figure 2 shows the comparison between gaussian width and reconstruction error, in which we observe that there is an approximate correlation between the two quantities. In particular, for sub-images 2,7 and 8—which mostly consist of background—both the gaussian width and the reconstruction error is small. On the other hand, images 3, 6 and 9, which consist of intricate patterns and objects, have large gaussian width as well as large reconstruction error. Consequently, we can deduce that for sub-images with large gaussian width, in order to achieve low reconstruction error, one would have consider a larger number of atoms m or larger sparsity k.

6. Conclusion and Open Questions

In this paper, we studied the problem of testing sparsity with respect to unknown and known bases. While the optimization variants of these problems (namely Dictionary Learning and Sparse Recovery) are known to be NP-hard in the worst case, our results show that under appropriate relaxations, these problems admit efficient property testing algorithms. Future work include designing testing algorithms for sparsity over an unknown basis with stronger guarantees or developing impossibility results. We also hope that this paper leads to study of property testing of other widely studied hypotheses in machine learning such as nonnegative rank and VC-dimension.

Acknowledgements

We would like to thank David Woodruff for showing us the sketching-based tester described in Section 1.2.

References

Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P., and Tandon, R. Learning sparsely used overcomplete dictionaries. In *Proc. 27th Annual ACM Workshop on Computational Learning Theory*, pp. 123–137, 2014.

Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. In *Proc. 28th Annual ACM Symposium on the Theory of Computing*, pp. 20–29. ACM, 1996.

Alon, N., Lee, T., Shraibman, A., and Vempala, S. The approximate rank of a matrix and its algorithmic applications: approximate rank. In *Proc. 45th Annual ACM Symposium on the Theory of Computing*, pp. 675–684. ACM, 2013.

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. Living on the edge: A geometric theory of phase transitions in convex optimization. *CoRR*, abs/1303.6672, 2013.

Arora, S., Ge, R., and Moitra, A. New algorithms for learning incoherent and overcomplete dictionaries. In

[6] For a matrix $\mathbf{Y} \in \mathbb{R}^{d \times n}$ approximated by overcomplete basis $\mathbf{A}$ and coefficient matrix $\mathbf{X}$, the reconstruction error is equal to $\|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2/(n.d)$.

Proc. 27th Annual ACM Workshop on Computational Learning Theory, pp. 779–806, 2014.

Babai, L., Fortnow, L., and Lund, C. Non-deterministic exponential time has two-prover interactive protocols. *Computational Complexity*, 1(1):3–40, 1991.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Blum, A., Har-Peled, S., and Raichel, B. Sparse approximation via generating point sets. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pp. 548–557. SIAM, 2016.

Blum, M., Luby, M., and Rubinfeld, R. Self-testing/correcting with applications to numerical problems. *J. Comp. Sys. Sci.*, 47:549–595, 1993. Earlier version in STOC'90.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255.

Bourgain, J., Dirksen, S., and Nelson, J. Toward a unified theory of sparse dimensionality reduction in euclidean space. *Geometric and Functional Analysis*, 25(4):1009–1088, 2015.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Chierichetti, F., Gollapudi, S., Kumar, R., Lattanzi, S., Panigrahy, R., and Woodruff, D. P. Algorithms for ℓ_p low-rank approximation. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 806–814, 2017.

Czumaj, A., Sohler, C., and Ziegler, M. Property testing in computational geometry. In *Proc. 8th European Symposium on Algorithms*, pp. 155–166. Springer, 2000.

Dasgupta, A., Kumar, R., and Sarlós, T. A sparse Johnson-Lindenstrauss transform. In *Proc. 42nd Annual ACM Symposium on the Theory of Computing*, pp. 341–350. ACM, 2010.

Elad, M. and Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

Foster, D., Karloff, H., and Thaler, J. Variable selection is hard. In *Conference on Learning Theory*, pp. 696–709, 2015.

Goldreich, O., Goldwasser, S., and Ron, D. Property testing and its connection to learning and approximation. *J. ACM*, 45:653–750, 1998.

Gordon, Y. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Kane, D. M. and Nelson, J. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):4, 2014.

Kane, D. M., Nelson, J., and Woodruff, D. P. An optimal algorithm for the distinct elements problem. In *Proc. 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems*, pp. 41–52. ACM, 2010.

Krauthgamer, R. and Sasson, O. Property testing of data dimensionality. In *Proc. 14th ACM-SIAM Symposium on Discrete Algorithms*, pp. 18–27. Society for Industrial and Applied Mathematics, 2003.

McGregor, A. Graph stream algorithms: a survey. *ACM SIGMOD Record*, 43(1):9–20, 2014.

Mendelson, S. and Vershynin, R. Entropy, combinatorial dimensions and random averages. In *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings*, pp. 14–28, 2002.

Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Razaviyayn, M., Tseng, H.-W., and Luo, Z.-Q. Dictionary learning for sparse representation: Complexity and algorithms. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5247–5251. IEEE, 2014.

Ron, D. Property testing: A learning theory perspective. *Foundations and Trends in Machine Learning*, 1(3):307–402, 2008.

Rubinfeld, R. and Shapira, A. Sublinear time algorithms. In *Proc. International Congress of Mathematicians 2006*, volume 3, pp. 1095–1110, 2006.

Rubinfeld, R. and Sudan, M. Robust characterizations of polynomials with applications to program testing. *SIAM J. on Comput.*, 25:252–271, 1996.

Rudelson, M. and Vershynin, R. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

Slepian, D. The one-sided barrier problem for gaussian noise. *The Bell System Technical Journal*, 41(2):463–501, 1962.

Spielman, D. A., Wang, H., and Wright, J. Exact recovery of sparsely-used dictionaries. In *Proc. 25th Annual ACM Workshop on Computational Learning Theory*, pp. 37–1, 2012.

Tillmann, A. M. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letters*, 22(1):45–49, 2015.

Vershynin, R. Lectures in geometric functional analysis. *Preprint, University of Michigan*, 2011.

Vershynin, R. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance*, pp. 3–66. Springer, 2015.

Vershynin, R. High dimensional probability, 2016.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.