Testing for Dictionary Learning(ness)

Md Enayat Ullah
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Kanpur, India
enayat@iitk.ac.in

Deepanshu Gupta
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Kanpur, India
dpanshu@iitk.ac.in

Abstract

Dictionary Learning (or Sparse Coding) is a problem wherein given a collection of vectors, the goal is to learn a dictionary/basis (which may be over-complete) in which the vectors have a sparse representation. Various constraints and assumptions like Restricted Isometric Property(RIP)/ Incoherence are required for the well-posedness of the problem. However, the objective of this project is not to find the dictionary or the sparse representations, but rather to construct deterministic property tests which given a set of vectors says whether there exist a dictionary under which the vectors have sparse representations or not. This in turn translates to crucially using the RIP property of the dictionary. In this work, we attempt to find a poly-time algorithm to test for dictionary learning.

1 Introduction

Dictionary Learning is a ubiquitous problem in signal processing, neuroscience, machine learning etc. The objective is, given to set of vectors, to recover a basis which allows for sparse realizations of given vectors ($Y = AX$). In the general case, the problem is ill-posed and NP-Hard. However, standard assumptions on the basis and the sparsity of representations make this problem computationally tractable. Since the problem subsumes the problem of sparse recovery, the assumption of Restricted Isometry Property (RIP) on the basis matrix is borrowed from Candes and Tao's seminal work on Compressive Sensing[CT06]. The dictionary is usually allowed to be over-complete as it provides greater flexibility for generating sparse representations. Moreover, we have upper bounds on the allowed sparsity of X for provable guarantees.

Dictionary learning is present in a plethora of day-to-day natural and man-made applications. Images exhibit sparse representations in the spectral basis. This is crucially used in high-speed photography wherein the camera stores the dictionary of images in the hardware and instead of storing the raw images, their sparse representations(in the stored dictionary) is kept. This leads to less memory requirement and processing. Moreover, it is also used for image de-blurring and digital zoom in cameras. Dictionary learning also finds its application in Brain Imaging.

Methods for Dictionary learning mostly comprise of heuristics like K-SVD, gradient descent style methods, Method of Optimal directions (MOD) etc. Recently, there are works which give provable guarantees with mostly standard set of constraints and assumptions mentioned above. Moreover, these works assume a generative model for X, either from a combination of Bernoulli-subgaussian, or sampling support from a discrete distribution followed by sampling the entries from subgaussian. [SWW12] analyzed this in the noiseless setting and when A is a basis. [AAJ$^+$14] and [AGM14] independently gave algorithms which work for the over-complete case. Moreover, [AGMM15] provided a neural framework which again provided provable guarantees. [AAN13] further proved the global optimality of an alternating minimization based approach given that it is initialized close to the true solution.

In this work, we are not concerned with the learning problem and our objective is not the recovery of A and X. We rather work on developing and analyzing methods which given a matrix Y outputs either a **YES** or **NO** answer, if the matrix allows a (column)sparse representation or if they are far away from such a sparse representation respectively.

Therefore we solve the property testing analogue/relaxation of the learning problem. Also, unlike the recovery related works which assume a generative model for data, we work in the agnostic setting.

2 Property Testing

Property testing is defined in [Ron08] as ”Given the ability to perform (local) queries concerning a particular object the problem is to determine whether the object has a predetermined (global) property or differs significantly from any object that has the property. In the latter case we say it is far from (having) the property. The algorithm is allowed a small probability of failure, and typically it inspects only a small part of the whole object”. Property testing and hypothesis testing in statistics literature are very related. A definition of property testing in probabilistic scenario is given as below:

Definition 1. (Standard Testing) [Ron08] A testing algorithm for property $\mathcal{P}$ of functions from domain X to range R) is given a distance parameter ϵ and query access to an unknown function. $f : X \rightarrow R$ and let $dist(f, \mathcal{P})$ be the distance measure between function f and property $\mathcal{P}$.

- If $f \in \mathcal{P}$ then the algorithm should accept with high probability
- If $dist(f, \mathcal{P}) > \epsilon$ the the algorithm should reject with high probability

Property testing literature generally discuss notions of Completeness and Soundness to establish guarantees of algorithms. The two are defined below:

Definition 2. (Completeness) An algorithm $\mathcal{A}$ is said to be complete if given that an object satisfies property $\mathcal{P}$, the algorithm outputs YES.

Definition 3. (Soundness) An algorithm $\mathcal{A}$ is said to be sound if given that an object is ϵ-”far” way from the property $\mathcal{P}$, the algorithm outputs NO.

3 Problem Formulation

Dictionary Learning problem can essentially be formalized as the following:

$$
\begin{aligned}
&Y \in \mathbb{R}^{d \times n} && (\textit{Input: } n \text{ samples of } p \text{ dimensional vectors}) \\
&A \in \mathbb{R}^{d \times m} && (\textit{Output: } \text{Dictionary}) \\
&X \in \mathbb{R}^{m \times n} && (\textit{Output: } \text{Sparse Representations})
\end{aligned}
$$

Given Y, find matrices A (basis/dictionary) and X such that

$$Y = AX, \text{ and } \|X_i\|_0 \leq k \;\forall\; i \text{ for a given } k \in [n]$$

Alternate way to look at the problem is: Let k' be defined as follows

$$k' = \max_i \min_{A, X_i \text{subject to} ||Y_i - AX_i|| \leq \epsilon} ||X_i||_0$$

Then, for a given value of ϵ say “Yes” if $k' \leq k$ and “No” otherwise.

4 Mathematical Preliminaries

4.1 Restricted Isometric Property

A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy *k-RIP* with RIP constant δ_k if $\forall\; X \in \mathbb{R}^n$ with $\|X\|_0 \leq k$, the following hold:

$$(1 - \delta_k) \|X\|_2 \leq \|AX\|_2 \leq (1 + \delta_k) \|X\|_2$$

Random matrices with iid entries and with sufficiently many measurements $\left(\Theta\left(\frac{kn\log n}{\delta_k^2}\right)\right)$ satisfy the RIP property with a high probability.

Theorem 1. [Preservation of angles between k-sparse vectors [HN07]] Suppose a matrix A satisfies RIP for k-sparse vectors with $\delta_k < \frac{1}{3}$. Then, for any vectors having sparsity at most k, supported on the same set of indices and separated by an angle $\alpha \in [0, \frac{\pi}{2}]$ and angle α_p between the projected vectors obey the following bound:

$$(1 - \sqrt{3\delta_k})\alpha \leq \alpha_p \leq (1 + \sqrt{3\delta_k})\alpha$$

Theorem 2. [Generalized RIP [HN07]] Under the condition of the above theorem inner product between two sparse vectors x and y, supported on the same set and separated by an acute angle α satisfies:

$$(1-\delta_k)||x||_2 \cdot ||y||_2 cos((1+3\delta_k)\alpha) \leq \langle Ax, Ay\rangle \leq (1+\delta_k)||x||_2 \cdot ||y||_2 cos((1-\sqrt{3\delta_k})\alpha)$$

4.2 Gaussian Width

The Gaussian Width of a set $\mathcal{S}$ is

$$\omega(S) = \mathop{\mathbb{E}}_{g}\left[\mathop{\mathbf{sup}}_{v\in\mathcal{S}}\langle g, v\rangle\right]$$

where $g \in \mathbb{R}^d$ is a random vector drawn from $\mathcal{N}_d\,(0, 1)$.

Lemma 3. The following holds true for Gaussian width $\omega\,(S)$ of a set $\mathcal{S}$, where C denotes a universal constant:

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

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

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

(iv) $\omega(\binom{d}{k}) \leq 2\sqrt{3k\log(d/k)}$ when $d/k > 2$ and $k \geq 4$.

4.3 Incoherence

The incoherence parameter μ of a matrix $A_{d\times m} = [a_1\; a_2\; \ldots\; a_m]$ is defined as

$$\frac{\mu}{\sqrt{d}} := \max_{r\neq s} |\langle a_r, a_s\rangle| \tag{1}$$

From here we shall consider the columns of matrix A to be normalized. In signal processing literature incoherence has often been called “mutual coherence” first introduced by [DH01]. Some other results associated with incoherence or mutual-coherence are:

- Let $A_{d\times m}$ be any matrix and μ be the incoherence paramter of A and $m > d$, then $\frac{\mu}{\sqrt{d}} \geq \sqrt{\frac{m-d}{d(m-1)}}$ [Wel74]
- Let $y \in \mathbb{R}^n$ and $y = Ax$ for some $x_{n\times 1}$ where x is k-sparse and A has incoherence parameter μ then x is unique sparsest vector for y if $k < \frac{\sqrt{n}}{\mu}$ [DET06, Fuc04]

Incoherence parameter has mostly found its use in analyzing the stability of solution of finding over-complete representation in signal processing, see [DET06, Fuc04].

5 Approach

We formalized approach that is very close to [BBG16], where in instead of using Gaussian width to give randomized tests for dictionary learning(by crucially using the RIP property of A), we are using incoherence parameter μ. Our objective has been to closely derive deterministic versions of the probabilistic theorems and results used by [BBG16] like Gordon’s theorem, Generalized Johnson-Lindenstrauss(GJL) lemma and Inverse Gordon lemma. Now, deterministic equivalence of Gordon’s theorem can be easily found using the incoherence parameter and we derived the equvalent Generalized Johnson-Lindenstrauss Lemma.

6 Results

In this section we shall give the completeness of our approach of property testing and also outline some ideas that we think can be used to approach soundness.

6.1 Completeness

Let $Y \in \mathbb{R}^{d\times p}$ such that $Y = A_{d\times m}X_{m\times p}$ with A being (δ, k) RIP and columns of X being k-sparse and normalized. Suppose, A is μ-incoherent then for all Y_i, Y_j columns of Y, $(1-\delta)^2 \leq |\langle Y_i, Y_j\rangle| \leq (1+\delta)^2$ or equivalently

$$\langle Y_i, Y_j\rangle \leq 1 + (k-1)\frac{\mu}{\sqrt{md}}$$

Proof: Proof of the first bound is trivial involving just the RIP property of A and k-sparsity of columns of X. We shall look into proving the incoherence equivalent condition. Let $Y = A_{d\times m}X_{m\times p}$ with above conditions satisfied and Y_i and Y_j be any two columns in Y. First, confirm that inner product of any two columns will be maximum when both vectors are same and weight is evenly distributed. Lower bound for the second part can be estimated using equivalence between upper bounds.

Now,

$$\begin{aligned}
|\langle Y_i, Y_j\rangle| &= |\langle \sum_{p=1}^{m} A_p X_{pi}, \sum_{q=1}^{m} A_q X_{qj}\rangle| \\
&\leq \sum_{p=1}^{m}\sum_{q=1}^{m} |\left\langle A_p, A_q \right\rangle||X_{pi}X_{qj}| \\
&\leq \sum_{p=1}^{m}\sum_{q=1}^{m} \frac{\mu}{\sqrt{d}}|X_{pi}X_{qj}| \\
&\leq 1 + (k-1)\frac{\mu}{\sqrt{md}}
\end{aligned}$$

Further, if X is incoherent with parameter μ_X, we have

$$\langle Y_i, Y_j\rangle \leq \frac{\mu_X}{\sqrt{m}} + (k-1)\frac{\mu}{\sqrt{md}}$$

6.2 Soundness

At this point we don't have a concrete soundness statement, however we do have some ideas that we are pursuing for soundness. Two of them are given below:

- Say that for a given Y corresponding A and X do not exist that satisfy isometry property or sparsity. Now, consider a special case such that there is a matrix with l-sparse columns and A matrix which is (δ_k, k)-RIP that satisfy $Y = AX$, here $l > k$. Now if all the elements in the l-sparse matrix are less than ω, then projection of that vector to k-sparse vector space should be $\mathcal{O}(\omega)$. But, if all elements of the l-sparse vector are greater than ω then the maximum inner product has to be greater than some function of l, ω, k. Hence, whenever the incoherence of Y is less than that function of l, ω, k, we can be sure that a A and X exist that are not $\mathcal{O}(\omega)$ far than our objective. Hence, we essentially assume a lower bound on the non-zero elements of sparse vectors which is a standard assumption in sparse recovery to prove support recovery of algorithms.
- Second idea is walk through the procedure of [BBG16] and find deterministic equivalents of the concepts used by them. This would involve finding a matrix that can transform any matrix to a space of sparse vectors. To this effort we derived an equivalent version of GJL lemma (given below). This would also require proving in the end that the transformation matrix will also preserve isometry on to the sparse vector space.

Lemma 4. Generalized Johnson-Lindenstrauss Lemma: Let $S \subset \mathcal{R}^n$. Then there exists a linear transformation $\Phi : \mathbb{R}^n \to \mathbb{R}^d$, where $d = \mathcal{O}\left(\frac{\mu\left(S\right)^2}{\epsilon^2}\right)$ and $\mu\left(S\right)$ is the incoherence parameter of the set S.

Proof:
Proof by Construction.

$$\text{We want } \left(1-\epsilon\right)\left\|X\right\|_2 \leq \left\|\Phi X\right\|_2 \leq \left(1+\epsilon\right)\left\|X\right\|_2 \text{ where } X \in S$$

We know

$$\min_{X\in\mathbb{R}^n}\left\|\Phi X\right\|_2 \leq \min_{X\in S}\left\|\Phi X\right\|_2 \leq \max_{X\in S}\left\|\Phi X\right\|_2 \leq \max_{X\in\mathbb{R}^n}\left\|\Phi X\right\|_2$$

Consider all vectors of S to be of unit norm, and $d \leq n$.

$$\min_{X\in\mathbb{R}^n,\left\|X\right\|_2=1}\left\|\Phi X\right\|_2 \leq \min_{X\in S}\left\|\Phi X\right\|_2 \leq \max_{X\in S}\left\|\Phi X\right\|_2 \leq \max_{X\in\mathbb{R}^n,\left\|X\right\|_2=1}\left\|\Phi X\right\|_2$$
$$\sigma_d^2 = \min_{X\in\mathcal{R}^n,\left\|X\right\|_2=1}\left\|\Phi X\right\|_2 \leq \max_{X\in\mathcal{R}^n,\left\|X\right\|_2=1}\left\|\Phi X\right\|_2 = \sigma_1^2$$

Here σ_1^2 and σ_d^2 are the largest and smallest singular values of Φ respectively.

We now construct a matrix $A \in \mathbb{R}^{n\times n}$ such that:

$$A_{ij} = \begin{cases} 1 & \text{if } i \neq j \\ \dfrac{\sqrt{n}\left\langle X_i, f_{ij}\left(\bar{X}_i\right)\right\rangle}{\sqrt{d}(n-1)} & \text{otherwise} \end{cases}$$

Where $f_{ij}\left(\bar{X}_i\right)$ is a convex combination of the vectors in S except X_i. We have

$$|f_{ij}\left(\bar{X}_i\right)| \leq \frac{\mu}{\sqrt{n}}$$
$$\sum_{j\neq i}\frac{\sqrt{n}\left\langle X_i, f_{ij}\left(\bar{X}_i\right)\right\rangle}{\sqrt{d}(n-1)} \leq \frac{\mu}{\sqrt{d}}$$

By Gershgorin's theorem, we have bounds on all the eigenvalues (λ_i) of A. [Ste75]

$$(1-\frac{\mu}{\sqrt{d}}) \leq \lambda_i \leq (1+\frac{\mu}{\sqrt{d}})\ \forall\ i \in [n]$$

Let Φ = SVD(A, d), rank-d thin SVD of A. Therefore

$$(1-\frac{\mu}{\sqrt{d}}) \leq \sigma_d^2 \leq \sigma_1^2 \leq (1+\frac{\mu}{\sqrt{d}})$$

Where σ_1^2 and σ_d^2 are the largest and smallest singular values of Φ respectively. Else if $d > n$, we construct the matrix $A \in \mathbb{R}^{d\times d}$, with the following entries:

$$A_{ij} = \begin{cases} 1 & \text{if } i \neq j \\ \dfrac{\sqrt{n}\left\langle X_i, f_{ij}\left(\bar{X}_i\right)\right\rangle}{\sqrt{d}(d-1)} & \text{otherwise} \end{cases}$$

The bounds derived on the singular values of Φ earlier follows in this case as well.
Therefore we have:

$$\sigma_d^2 = \min_{X\in\mathbb{R}^n,\left\|X\right\|_2=1}\left\|\Phi X\right\|_2 \leq \max_{X\in\mathbb{R}^n,\left\|X\right\|_2=1}\left\|\Phi X\right\|_2 = \sigma_1^2$$
$$(1-\frac{\mu}{\sqrt{d}}) \leq \left\|\Phi X\right\|_2 \leq (1+\frac{\mu}{\sqrt{d}}),\ \forall\ X \in S$$

Define $\epsilon = \dfrac{\mu}{\sqrt{d}}$, we get $d = \dfrac{\mu^2}{\epsilon^2}$. Hence we have a matrix Φ such that it is an ϵ-isometry on S. The existence thus essentially follows from the construction of such a matrix Φ.

The proof for soundness in [BBG16] essentially involved two steps: Projection and Covering

- **Projection:** Sets with small Gaussian width point sets can be almost isometrically embedded into a low dimensional subspace.
- **Covering:** Appropriately sparse point sets on projection to a low dimensional subspace form a cover of the unit sphere on the smaller dimension

We have proved the analogue of projection above with the Gaussian Width replaced by Incoherence (Generalized Johnson-Lindenstrauss Lemma), and it appears that the subsequent step Covering does not involve the use of Gaussian Width. However we won't explicitly state now that their results extend naturally to our case (even if there is no Gaussian width involved, their randomized proofs might hinder with our hopefully deterministic guarantees). We want to go reproduce their proof more clearly and see if it applies to our case, and it it does, what is the final soundness statement.

7 Experiments

We ran experiments to test our upper bound on incoherence, which validates our claim of deterministically testing the completeness case. We generated the matrices from standard multivariate normal distributions, and the number of rows we sample guarantees RIP with high probability. The support of X is sampled from a uniform distribution followed by a ceiling operation, and the entries further sampled from standard normal. Figure 1 compares the incoherence of A and Y, as well as our upper bound on the Incoherence of Y. It is apparent the the upper bound is considerably tight, as the calculated Incoherence of Y just underestimates the bound. Figure 2 contain the parameters of Figure 1 along with Gaussian Width. This shows the the Gaussian width behaves very similar to Incoherence.

In order to empirically observe how Incoherence and Gaussian width relate to each other, we investigated how they behaves with respect to the vector dimensionality and the cardinality of the set. Note that the experiments are performed on matrices whose entries are sampled from standard normal, and the performance averaged over a large number of experiments. Therefore these essentially capture the expected behaviour with respect to the distribution used to generate the matrix. Figure 3 and 4 show how Incoherence and Gaussian Width behave with respect to vector dimensionality and set cardinality respectively.

8 Conclusion

In this project, we commenced work on deterministic tests for Dictionary Learning. This work is extensively inspired from Barman et al.'s [BBG16] work wherein they give randomized test for the same problem. We established the completeness of the property testing routine given that the RIP parameters of the dictionary generating the samples is known to us. We established an upper bound on the incoherence of Y based on those parameters. We also ran matlab simulations to validate our completeness claim. Our direction for the soundness guarantee is nascent and exploratory and we aim to work on it to get succinct results.

Our result currently says that when the incoherence is greater than the threshold, no dictionary learning on the parameters specified is possible. However, when the value is below, we essentially give the cases which cannot happen under the parameters specified. We further aim to expand this terrain of ruled-out cases so as to give a good enough result in the yes case. One line of work we aim to explore is if we assume that X is incoherent, either by explicitly assuming the incoherence parameter or by assuming a generative model for X. In the generative model case, we get an expected incoherence parameter of X, which effects the incoherence of Y substantially. However this would make out testing routine probabilistic as opposed to the current deterministic algorithm. Moreover,another assumption we want to make is that the non-zero entries of X are lower bounded by a certain quantity. This assumption is usually made in sparse recovery to show support recovery. We hope that under either of these assumptions, our soundness guarantees will become more clear and we eventually give a polytime deterministic test for Dictionary Learning.

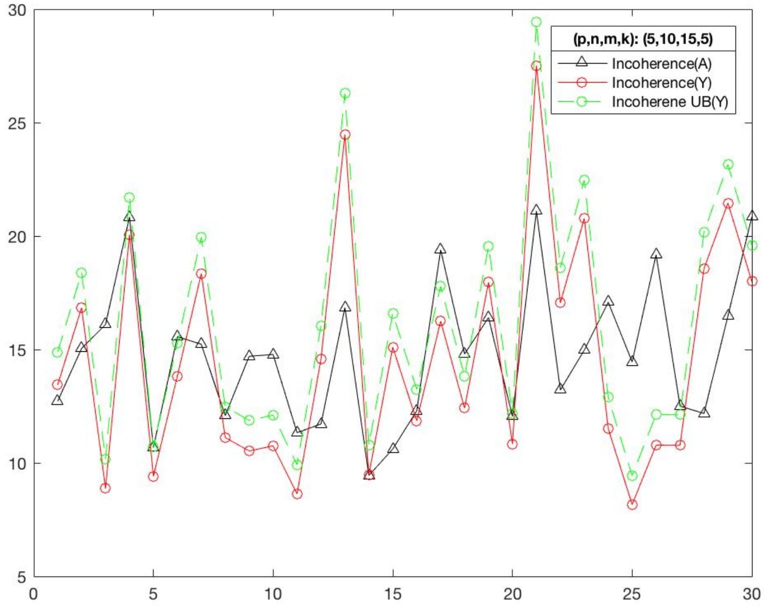

Figure 1: Incoherence of Y, A and upper-bound

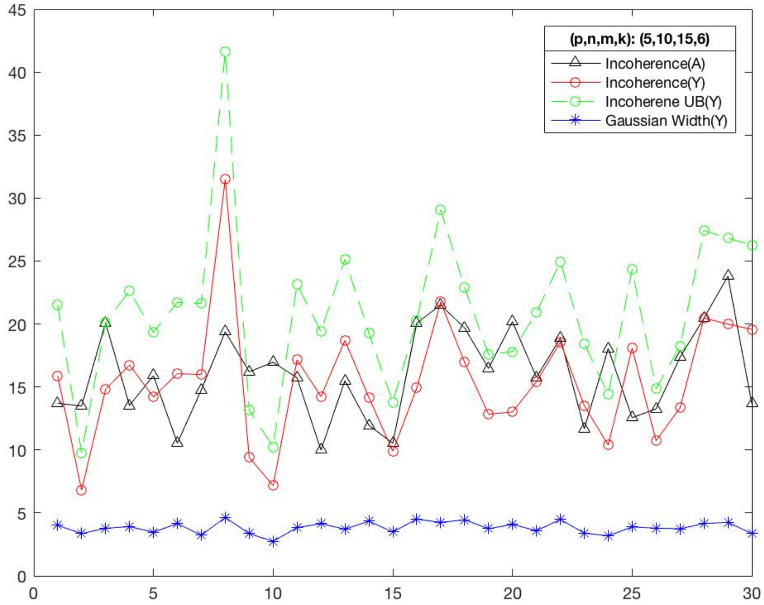

Figure 2: Incoherence of Y, A,upper-bound and Gaussian Width

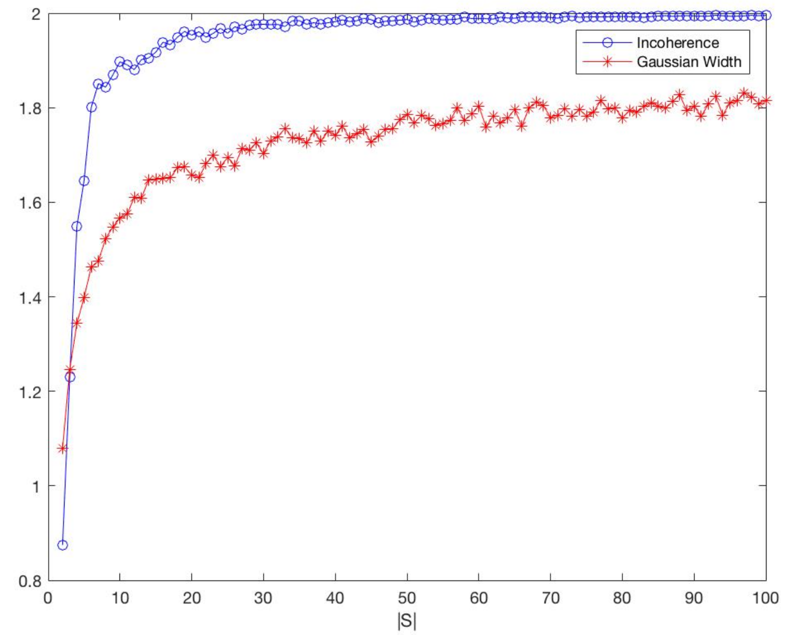

Figure 3: Incoherence and Gaussian Width VS. Set Cardinality

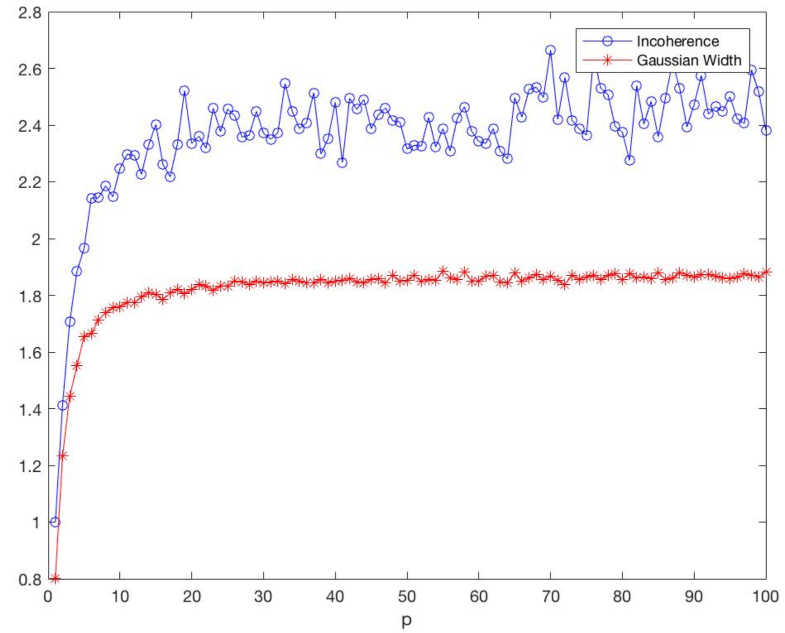

Figure 4: Incoherence and Gaussian Width VS. Vector Dimensionality

References

[AAJ+14] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *COLT*, pages 123–137, 2014.

[AAN13] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *stat*, 1050:8, 2013.

[AGM14] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *COLT*, pages 779–806, 2014.

[AGMM15] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.

[BBG16] Siddharth Barman, Arnab Bhattacharyya, and Suprovat Ghoshal. The dictionary testing problem. *arXiv preprint arXiv:1608.01275*, 2016.

[CT06] 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.

[DET06] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2006.

[DH01] David L Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[Fuc04] J-J Fuchs. On sparse representations in arbitrary redundant bases. *IEEE transactions on Information theory*, 50(6):1341–1344, 2004.

[HN07] Jarvis Haupt and Robert Nowak. A generalized restricted isometry property. *University of Wisconsin-Madison, Tech. Rep. ECE-07-1*, 2007.

[Ron08] Dana Ron. Property testing: A learning theory perspective. *Foundations and Trends® in Machine Learning*, 1(3):307–402, 2008.

[Ste75] Gilbert W Stewart. Gershgorin theory for the generalized eigenvalue problem = . *Mathematics of Computation*, 29(130):600–606, 1975.

[SWW12] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *COLT*, pages 37–1, 2012.

[Wel74] Lloyd Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, 20(3):397–399, 1974.