Testing Sparsity

Arnab Bhattacharyya
Indian Institute of Science, Bangalore

Siddharth Barman

Suprovat Ghoshal

Learning versus Testing

Learning: Given examples $(x, f(x))$ where f is an unknown function in hypothesis class $\mathcal{H}$, find approximation of f w.r.t. a norm.

Testing: Given examples $(x, f(x))$, is f from the hypothesis class $\mathcal{H}$, or is f far from $\mathcal{H}$ w.r.t. a norm?

Property Testing

- Prelude to learning

- Typically based on a robust local characterization of membership in $\mathcal{H}$.

Extensively studied since ‘90’s for algebraic properties (linearity, membership in error-correcting codes, etc), graph-theoretic properties (bipartiteness, triangle-freeness, etc), expressibility as Boolean formulae, etc. **[Rubinfeld-Shapira ‘06, Ron ’08]**.

Testing Sparsity

Given a set of vectors $y_1, y_2, \ldots, y_p \in \mathbb{R}^d$, which of the two is true?

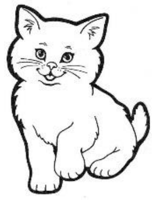
i. [**Structure**] There exists a matrix $A \in \mathbb{R}^{d \times m}$ and k-sparse vectors $x_1, \ldots, x_p \in \mathbb{R}^m$ such that $y_i \approx A x_i$ for all $i \in [p]$

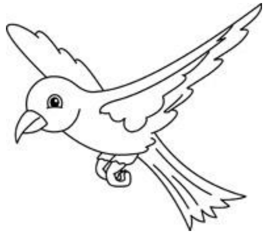
ii. [**Noise**] For every dictionary $A \in \mathbb{R}^{d \times m}$ and k-sparse vectors $x_1, \ldots, x_p \in \mathbb{R}^m$, $(y_1, \ldots, y_p)$ is "far" from $(Ax_1, \ldots, Ax_p)$

Property testing of a continuous property of real vectors

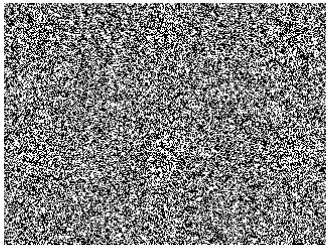
Motivation

Is there a different characterization of natural inputs?





versus



Sparse Coding & The Brain

Originally developed to explain early visual processing in the brain (edge detection) **[Olshausen-Field '96]**

Task: Given a set of image patches $y_1, \ldots, y_p$, learn a dictionary of bases $[\Phi_1, \Phi_2, \ldots, \Phi_m]$ minimizing both:

$$\sum_i ||y_i - \sum_j a_{ij} \Phi_j ||^2$$

and number of nonzero a_{ij}

Other applications

- Similar experiments done for early auditory processing and early somatosensory processing

- Widely used in machine learning now to learn natural feature representations for data

- Hierarchical sparse coding → Deep learning

Dictionary Learning Problem

Given a set of vectors $y_1, y_2, \dots, y_p \in \mathbb{R}^n$, find a matrix $A \in \mathbb{R}^{n \times m}$ and k-sparse vectors $x_1, \dots, x_p \in \mathbb{R}^m$ such that:

$$y_i \approx Ax_i \text{ for all } i \in [p]$$

- Considered a solved problem in practice: alternating minimization, K-SVD, etc
- For rigorous proofs, we need to make some assumption on the **dictionary** A and distribution of the inputs **[Spielman-Wang-Wright, Agarwal-Anandkumar-Jain-Netrapalli-Tandon, Arora et al]**. Somewhat unsatisfactory.

Motivation: Recap

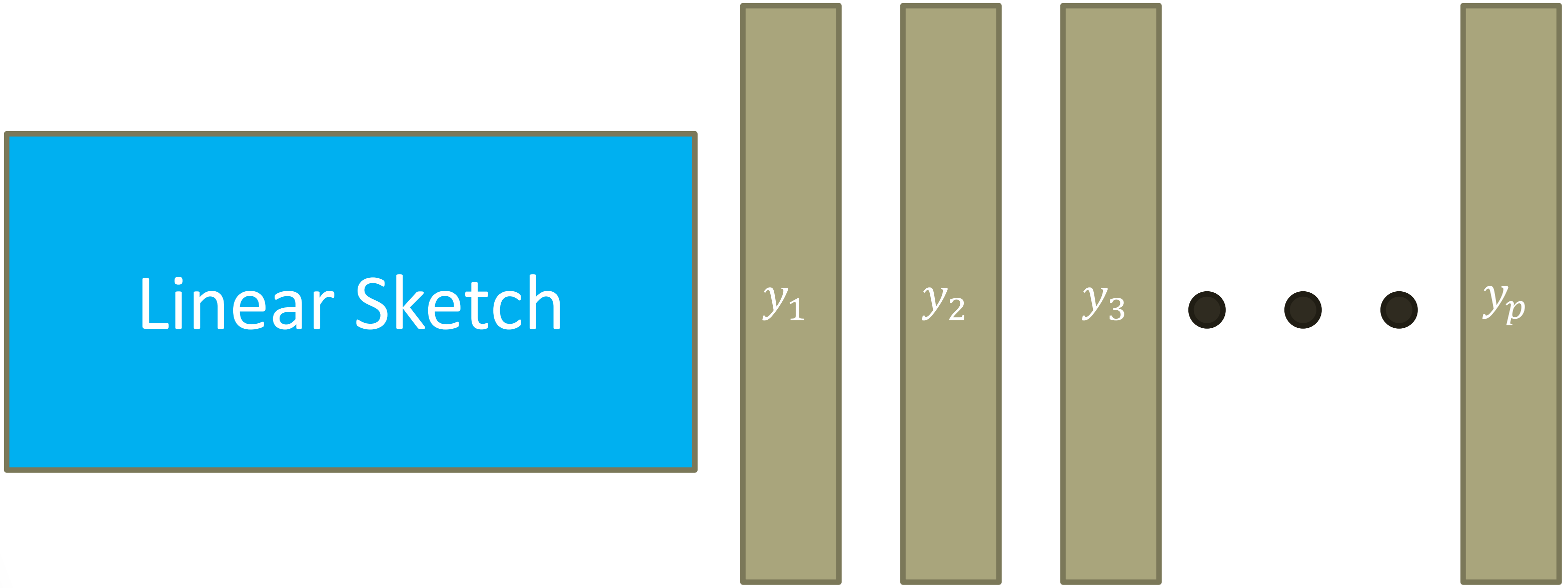
Is there a data-driven way to estimate the sparsity that's more efficient than learning the sparse representation?

- Could be useful in a scenario where most of the dataset is noise

A robust characterization of sparsity according to an unknown dictionary

Computational Model

- Linear measurements of input vectors



- **Query complexity**: # of rows in sketch matrix

Our Contribution

- Makes a connection between sparsity and **high-dimensional geometry**

- Algorithm estimates the ***gaussian width*** of the input vectors by projecting them into a constant-dimensional space

Dictionary Structure

- **RIP Assumption**: We assume that in the structured case, any submatrix of the dictionary matrix A with at most k columns is well-conditioned.

- Very common assumption in rigorous theorems about compressed sensing, sparse regression, sparse coding.

Main Theorem

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 $\mathbf{Y} = \mathbf{AX}$, 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*

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

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

Gaussian Width

Given a set $S \subseteq \mathbb{R}^n$:

$$\omega(S) = \mathbb{E}_g[\sup_{v \in S} \langle v, g \rangle]$$

where $g \in \mathbb{R}^n$ is a random Gaussian.

Our Tester

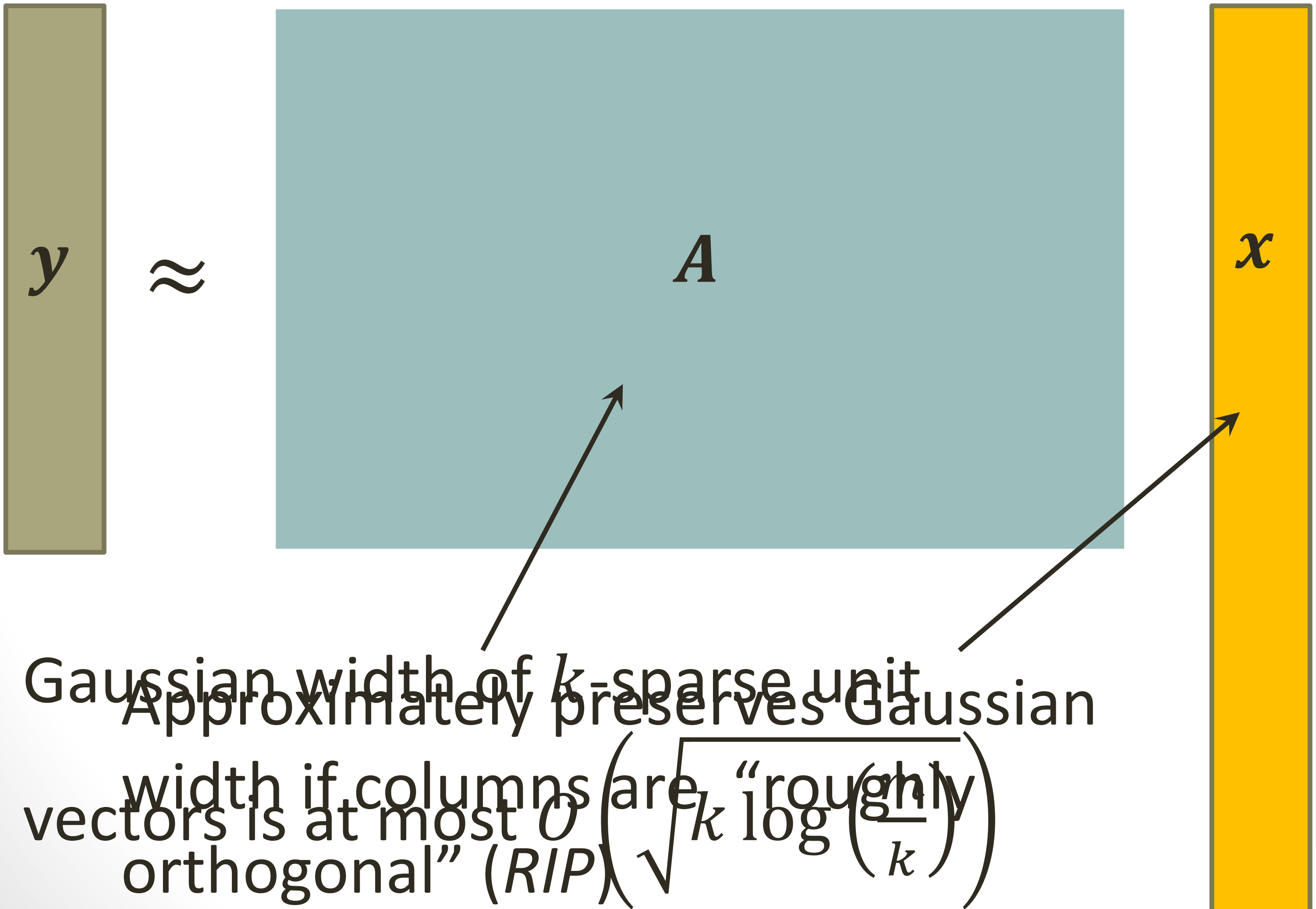
Estimate the Gaussian width by choosing a random Gaussian vector g and measure its correlation with all given vectors $y_1, \ldots y_p$. Accept if the estimated width is at most

$$\sim\sqrt{k \log\left(\frac{m}{k}\right)}.$$

Bounds on Width

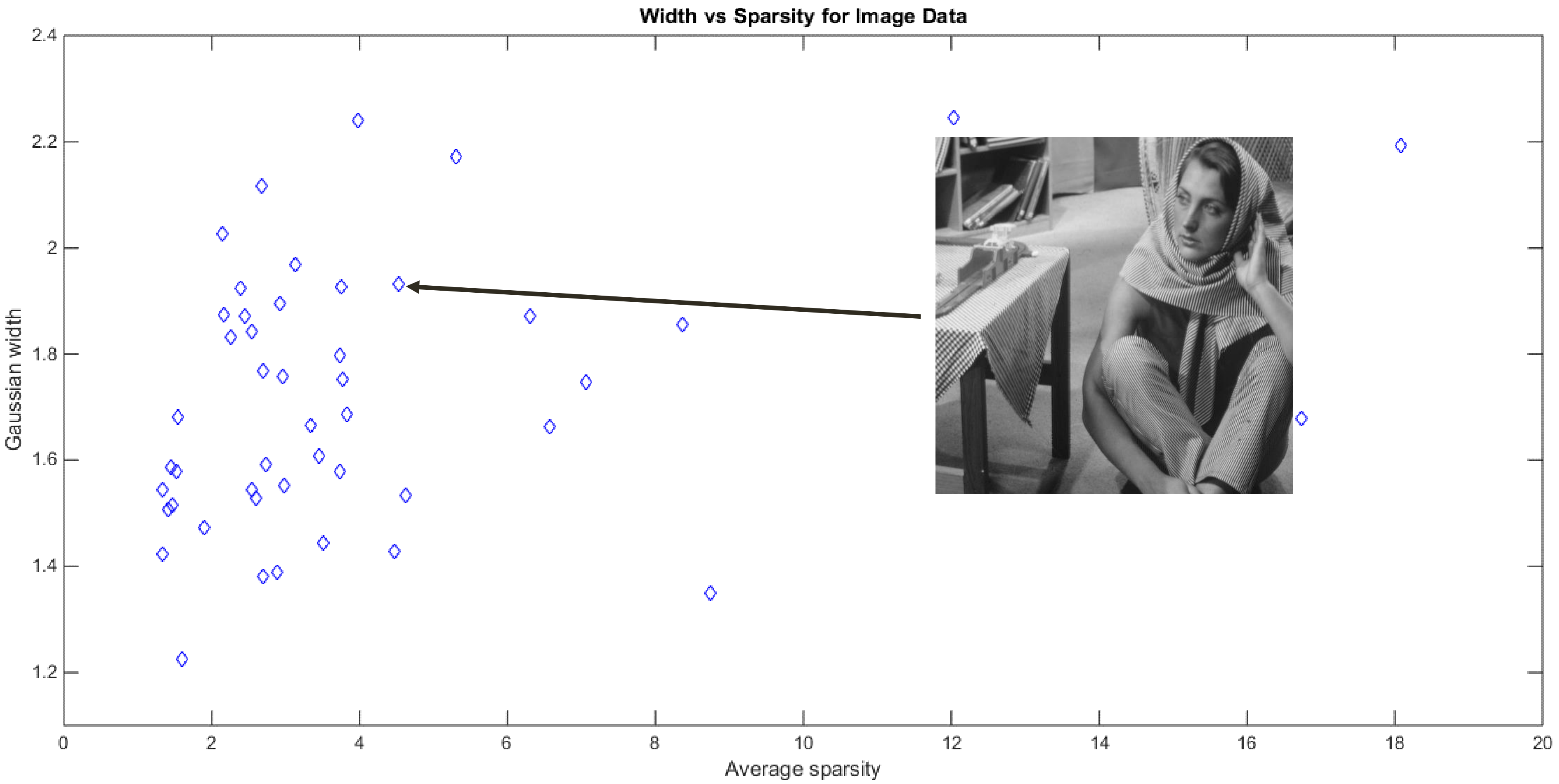
- If S is finite, $\omega(S) \lesssim \sqrt{\log |S|}$
- If S is of dimension k, then $\omega(S) \lesssim \sqrt{k}$
- If $S \subseteq \mathbb{R}^d$ consists of k-sparse vectors, then $\omega(S) \lesssim \sqrt{k \log\left(\frac{d}{k}\right)}$

Incoherent dictionaries



Gaussian width of k-sparse unit vectors is at most $O\left(\sqrt{k \log\left(\frac{n}{k}\right)}\right)$

Approximately preserves Gaussian width if columns are “roughly orthogonal” (*RIP*)



Width vs Sparsity for Image Data
Gaussian width
Average sparsity
2.4
2.2
2
1.8
1.6
1.4
1.2
0
2
4
6
8
10
12
14
16
18
20

Soundness

- But does the tester reject when the input is “far” from being sparsely coded?
- Equivalently, can we conclude approximate sparse coding when the Gaussian width is small?

Dimensionality

- $\omega^2(S)$ is a robust measure of "**intrinsic dimensionality**" of a data set.

- **Generalized Johnson-Lindenstrauss Theorem**: For any set $S \subseteq \mathbb{R}^d$, there is a linear map $\Phi: \mathbb{R}^d \to \mathbb{R}^n$ where $n = O\left(\frac{\omega^2(S)}{\epsilon^2}\right)$ such that Φ is an ϵ-isometry on S (preserves pairwise distances upto $1 \pm \epsilon$ factor)

Soundness Analysis

- Assume that $S = \{y_1, \dots, y_p\}$ have Gaussian width $< \sqrt{k \log\left(\frac{m}{k}\right)}$

- Will show that S is "close" to an incoherent linear map applied to $\Theta(k)$-sparse vectors in m dimensions.

Analysis Outline

Case 1

- Low Intrinsic Dimension

Case 2

- High Intrinsic Dimension

Case 1: $\omega(S) \lesssim \epsilon\sqrt{d}$

Lemma: With probability at least ½, for a uniformly chosen random rotation $R \sim \mathbb{O}_d$:

$$\max_{y \in R(S)} \|y\|_\infty \leq O\left(\frac{\omega(S)}{\sqrt{d}}\right)$$

So, in this case, $Y = RZ$, where R is a rotation and all entries of Z at most ϵ.

Case 2: $\omega(S) \gtrsim \epsilon\sqrt{d}$

- In this case, $d \leq O(k\epsilon^{-2}\log(m/k))$

Key Lemma: If $d \leq O(k\epsilon^{-2}\log(m/k))$, and $\Phi \sim \mathbb{R}^{d\times m}$ random gaussian matrix, then whp, $\Phi(S_\ell)$ is an $O(\epsilon^{1/4})$-cover of the unit sphere in d dimensions (after normalization). S_ℓ is the set of all $O(k\epsilon^{-4})$-sparse vectors in m dimensions.

- Hence, there exists set X of $O(k\epsilon^{-4})$-sparse vectors such that $\|y_i - \Phi(x_i)\| \leq O(\epsilon^{1/4})$.

Proof of Key Lemma

- **Gaussian width** strikes again!
- Informally, if a set of unit vectors $T \subset \mathbb{R}^n$ has gaussian width at least $\sqrt{n}(1-\epsilon)$, then for any unit vector x, whp over random rotations R, there is an element of $R(T)$ that is $O(\epsilon^{1/4})$ close to x in ℓ_2-norm.
- Proof uses this fact along with **lower bound** on gaussian width of ℓ-sparse vectors.

In Summary

We obtain a fast and robust distinguisher between sparse and very non-sparse sets of vectors (with respect to unknown dictionary).

Robust geometric characterization of sparse coding

Characterizations for other hypotheses classes in machine learning? Neural networks with 1 hidden layer?

Thanks!