Chaining, Random Matrices, and Dictionary Learning

Kyle Luh (Joint Work with Van Vu)

Department of Mathematics, Yale

June 23, 2016

Table of contents

Recovery Problem

- Let A be an $n \times n$ matrix, X be an $n \times p$ matrix.
- $$AX = Y \tag{1}$$

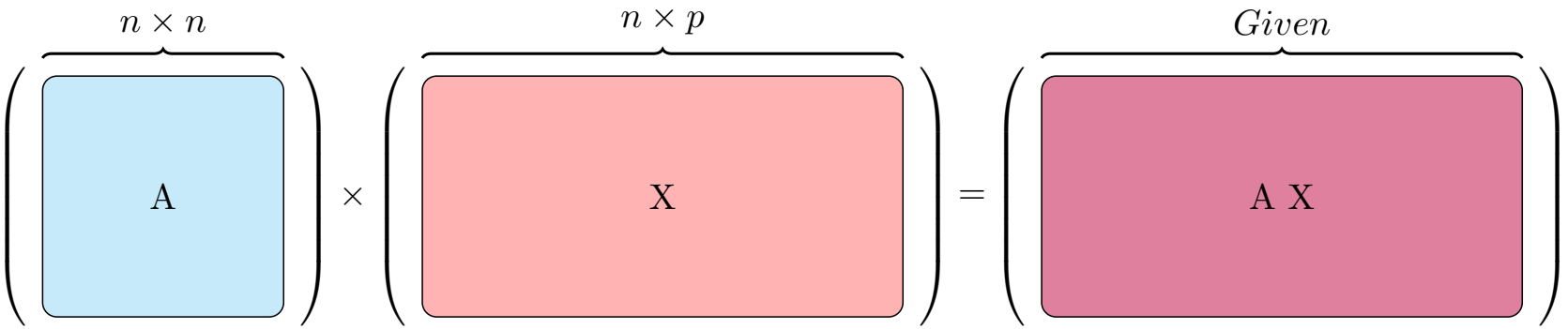


Figure: Recovery Problem

- Goal: find A and X, given Y.

Sparse Recovery

- In practice, X is frequently a sparse matrix.
- If X is sparse, the number of unknowns decreases dramatically, as the majority of entries of X are zero.
- The name of the game here is to find the minimum value of p, the number of observations, which guarantees a unique recovery.

Applications of Dictionary Learning

Learning dictionaries can be used for

- compression
- denoising
- classification
- recommendation systems
- blind source separation

Image Processing by Dictionary Learning

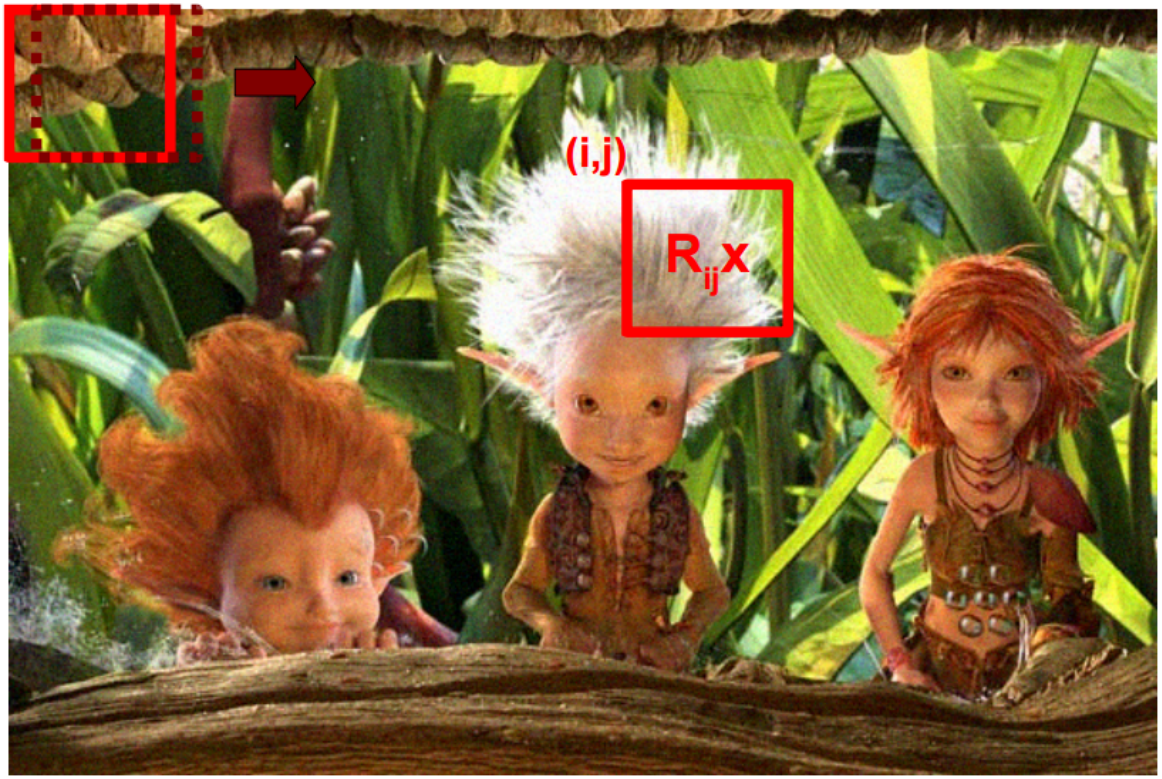


Figure: Source: Denoising and inpainting via dictionary learning by Tartavel et al

Image Processing by Dictionary Learning

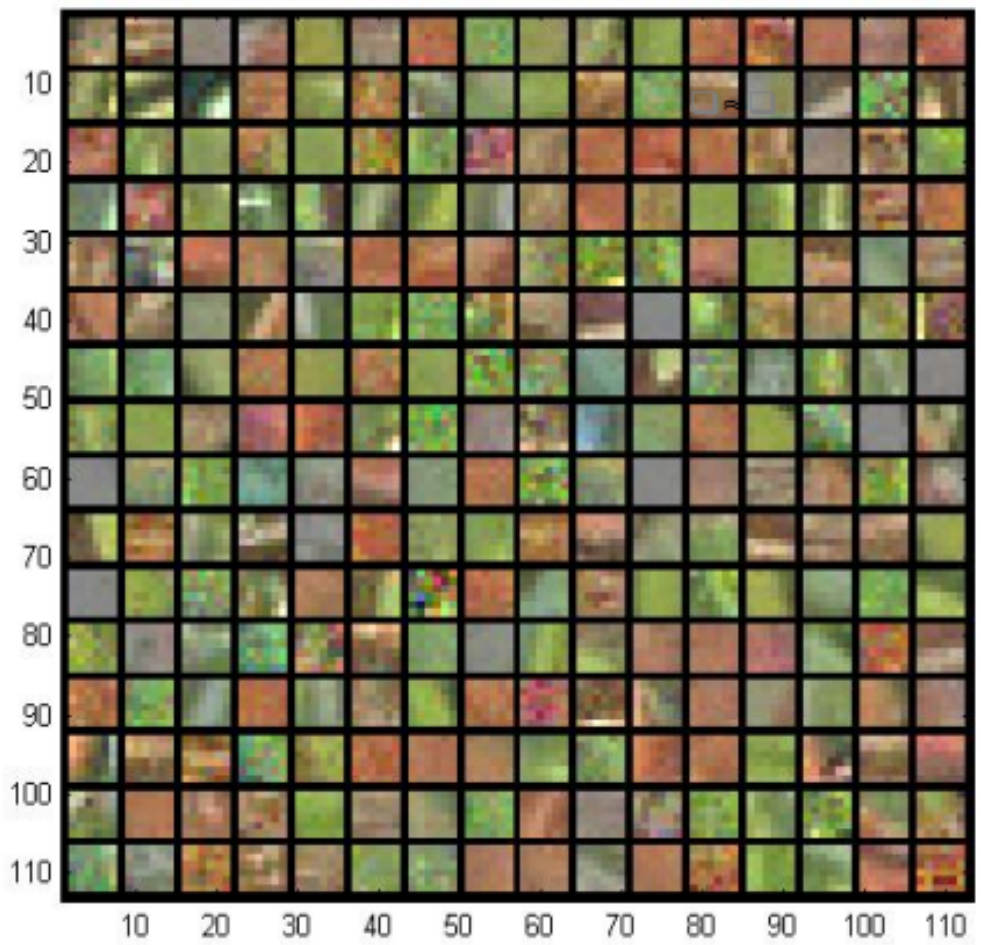
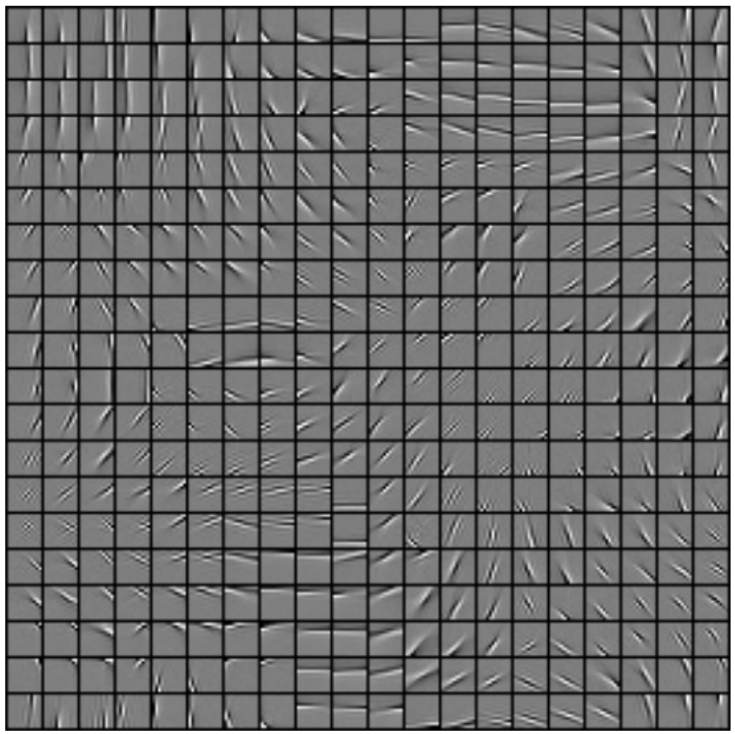
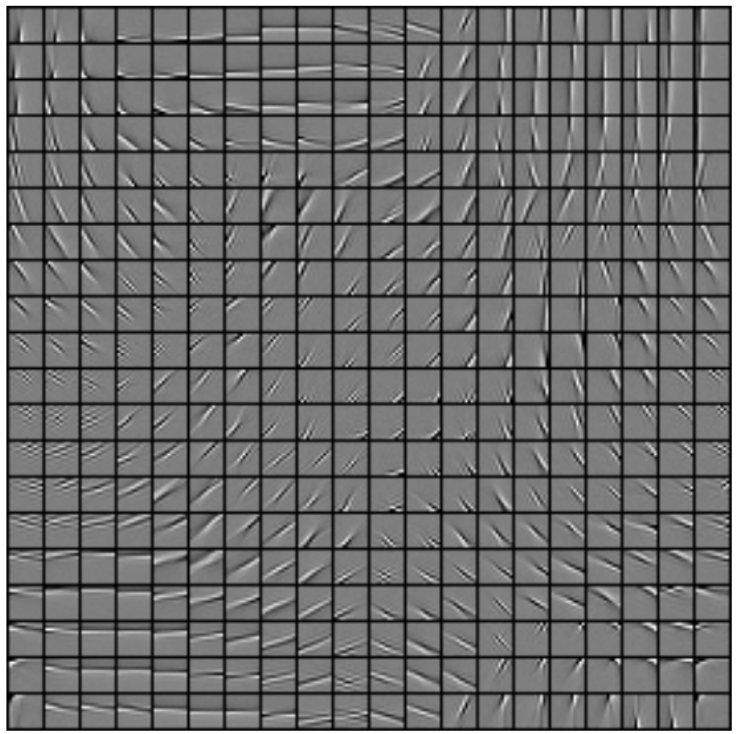


Image Denoising by Dictionary Learning



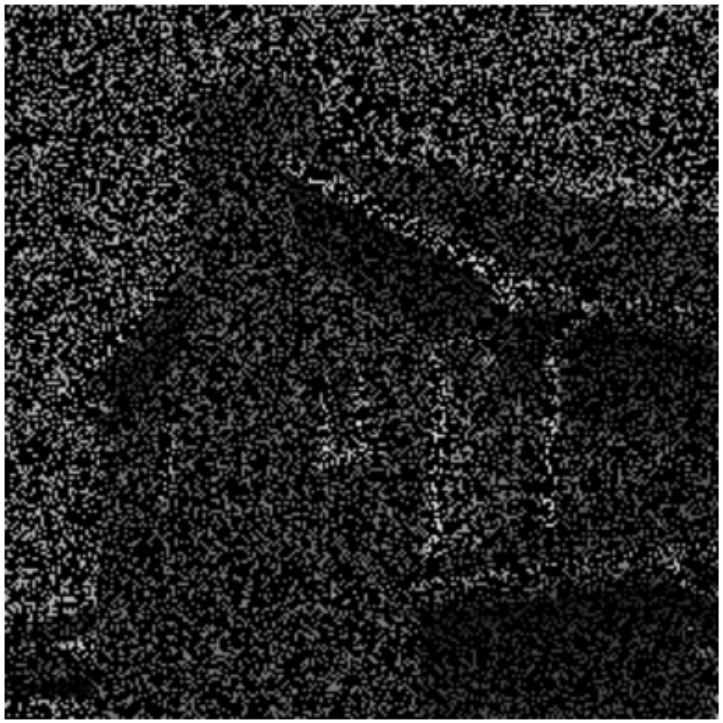
(a) With 3×3 neighborhoods.



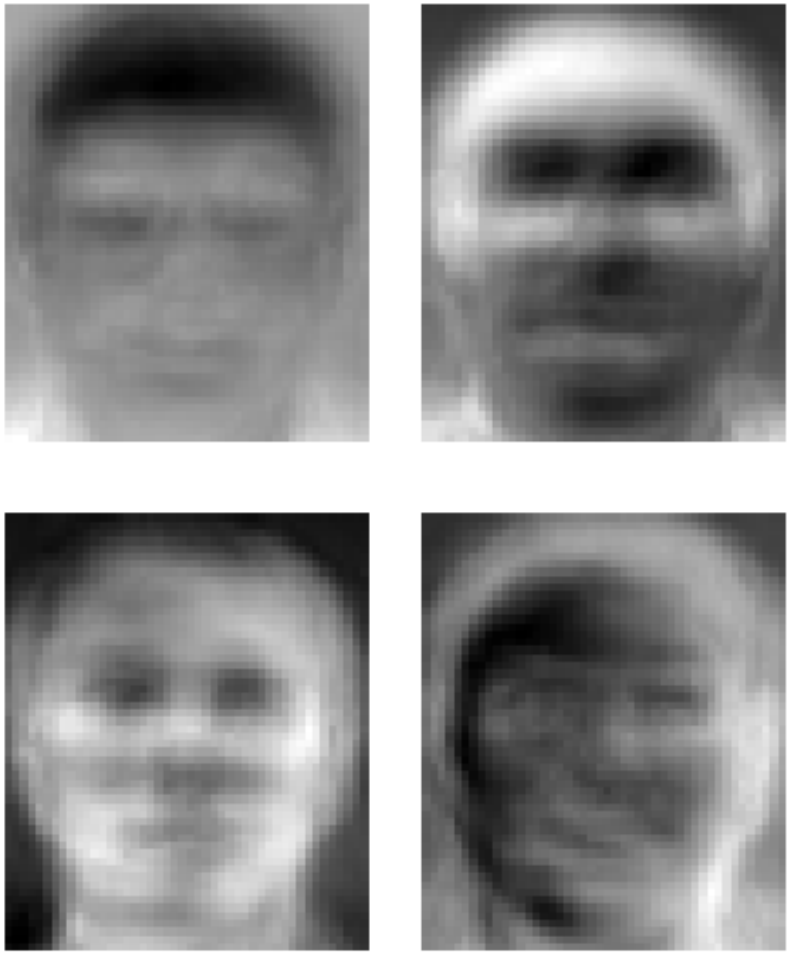
(b) With 4×4 neighborhood.

Image Denoising by Dictionary Learning

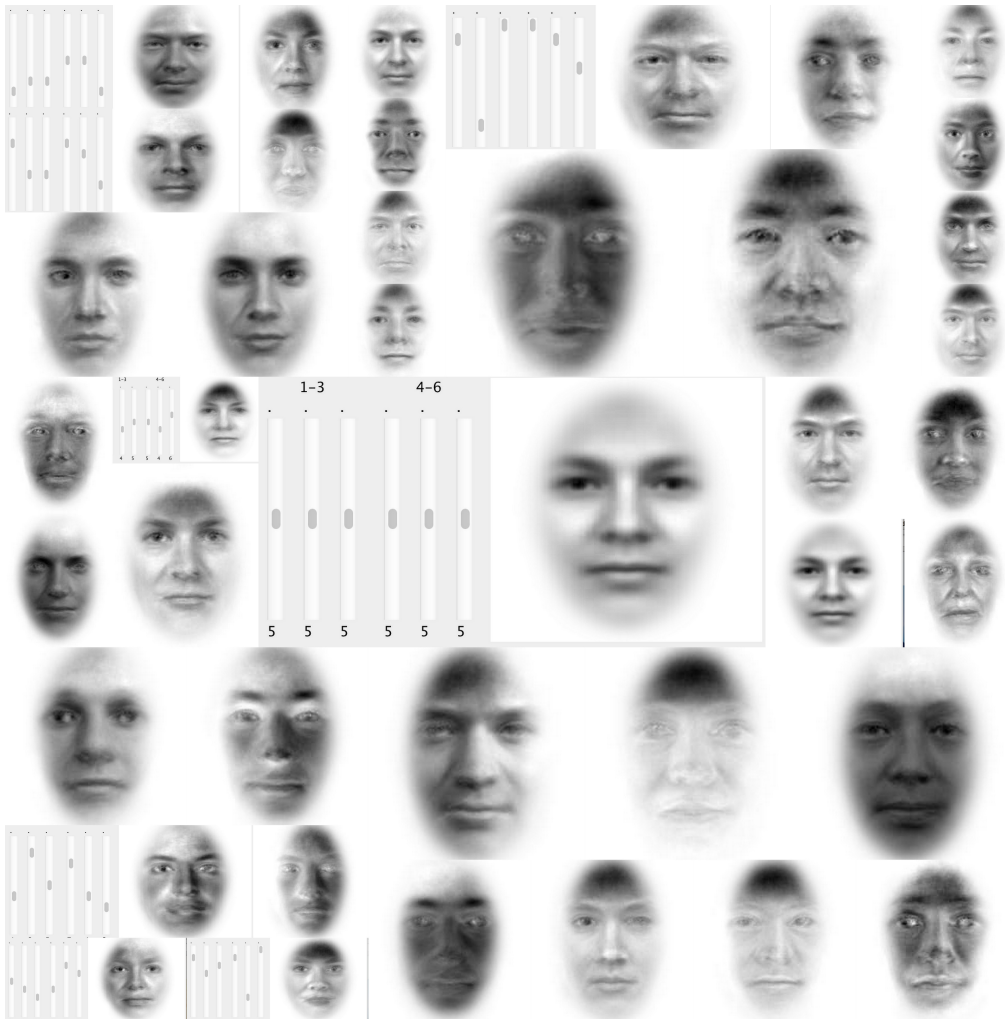
[Mairal, Sapiro, and Elad, 2008d]



Facial Recognition



Facial Recognition



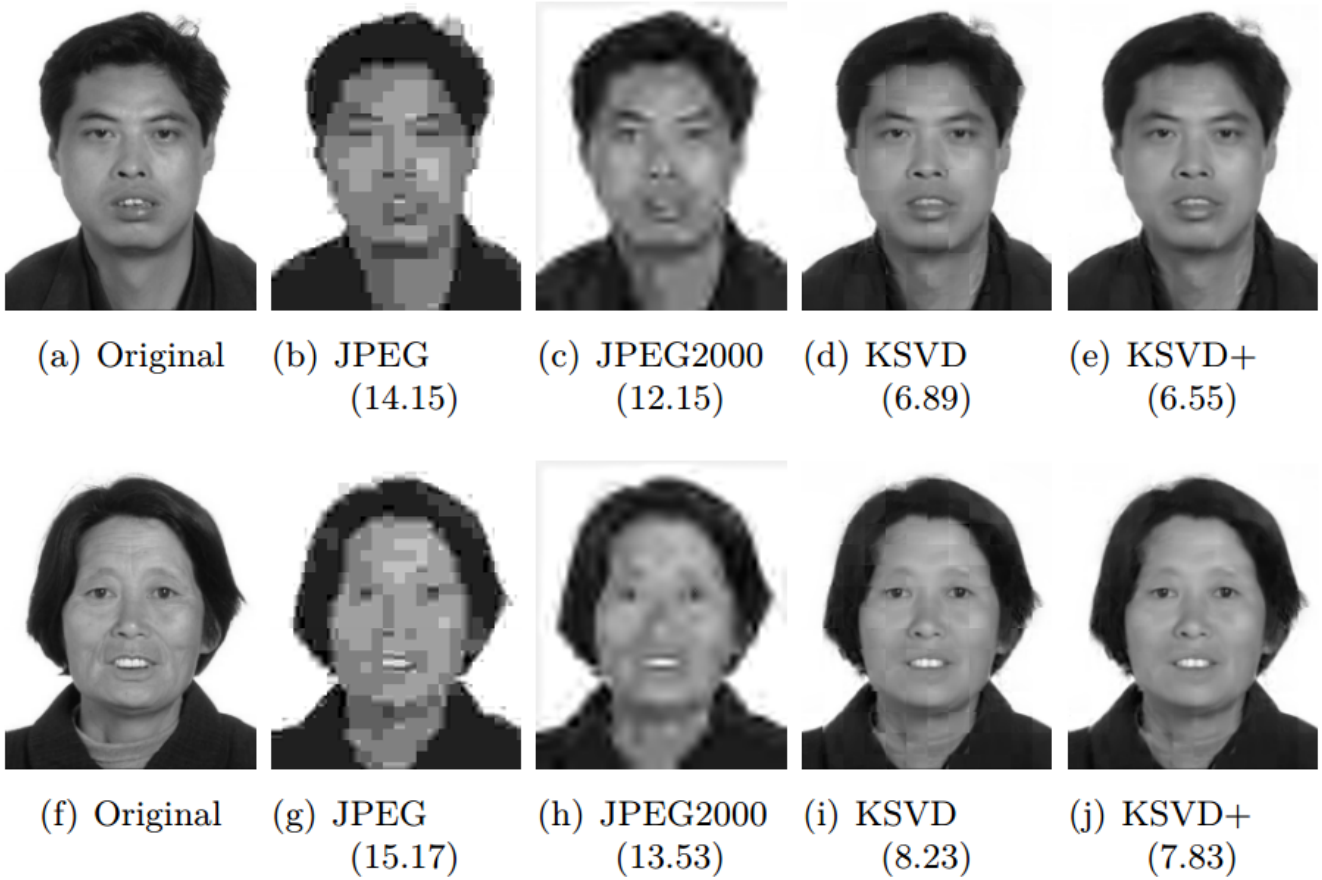
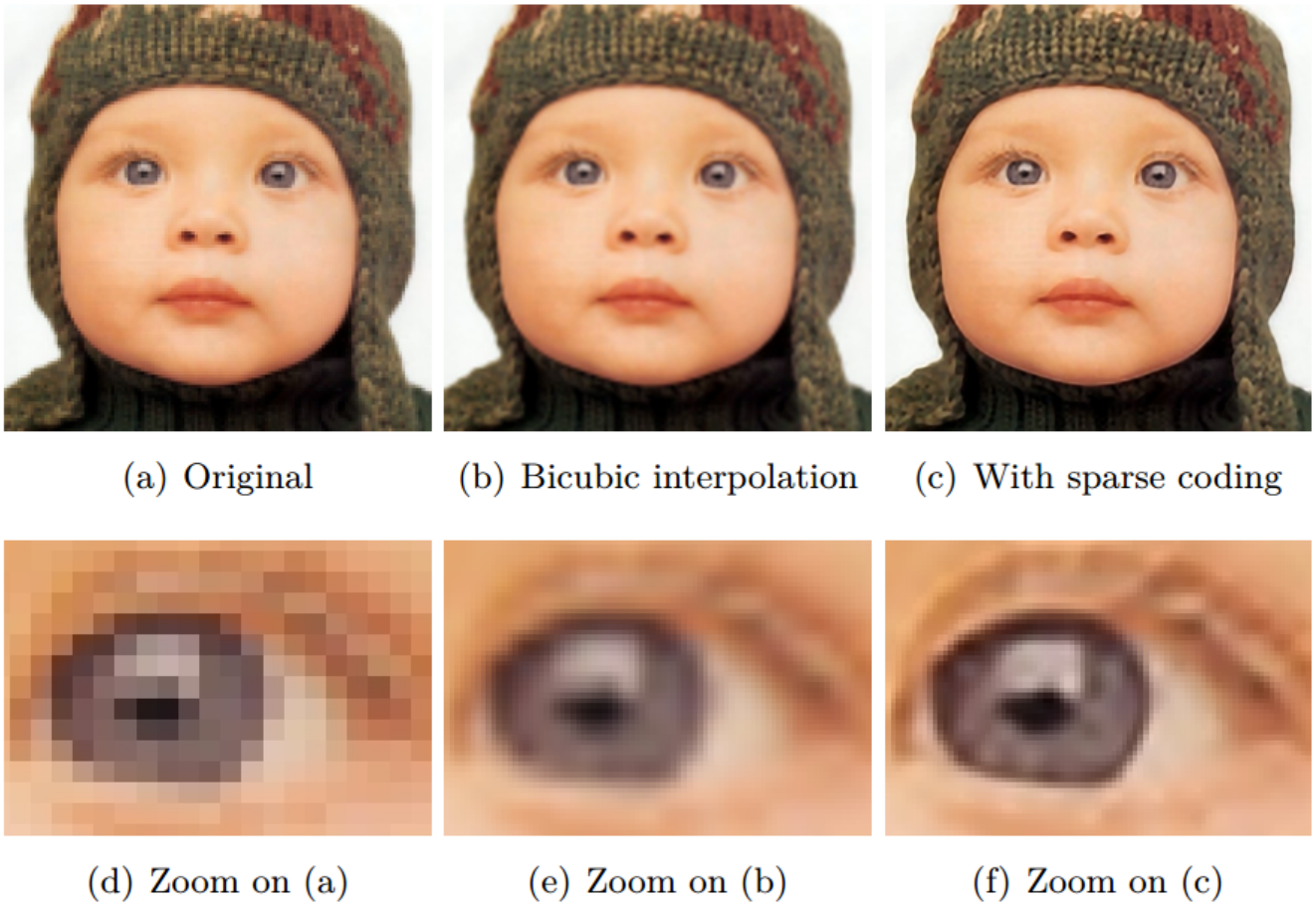


Figure: Facial Compression [2]

(a) Example A, Damaged

(b) Example A, Restored

Figure: Inpainting [2]



(a) Original (b) Bicubic interpolation (c) With sparse coding

(d) Zoom on (a) (e) Zoom on (b) (f) Zoom on (c)

Figure: Up-Scaling [2]

Two Main Problems

Problem 1. *What is the minimum value of p which guarantees a unique recovery in the theoretical sense ?*

Problem 2. *What is the minimum value of p which guarantees a unique recovery in the practical sense (with an efficient algorithm)?*

A random model and previous results

- X is a random Bernoulli-subgaussian matrix, defined as follows:

A random model and previous results

- X is a random Bernoulli-subgaussian matrix, defined as follows:
  - $$x_{ij} := \chi_{ij}\xi_{ij} \tag{2}$$

A random model and previous results

- X is a random Bernoulli-subgaussian matrix, defined as follows:
  - $$x_{ij} := \chi_{ij}\xi_{ij} \tag{2}$$
  - χ_{ij} are iid indicator random variables with $\mathbb{P}(\chi_{ij} = 1) = \theta$

A random model and previous results

- X is a random Bernoulli-subgaussian matrix, defined as follows:
  - $$x_{ij} := \chi_{ij}\xi_{ij} \tag{2}$$
  - χ_{ij} are iid indicator random variables with $\mathbb{P}(\chi_{ij} = 1) = \theta$
  - ξ_{ij} are iid, mean 0, variance bounded by 1

A random model and previous results

- X is a random Bernoulli-subgaussian matrix, defined as follows:
  - $$x_{ij} := \chi_{ij}\xi_{ij} \tag{2}$$
  - χ_{ij} are iid indicator random variables with $\mathbb{P}(\chi_{ij} = 1) = \theta$
  - ξ_{ij} are iid, mean 0, variance bounded by 1
  - $$\mathbb{P}(|\xi| > t) < 2\exp(-t^2/2).$$

A random model and previous results

- X is a random Bernoulli-subgaussian matrix, defined as follows:
  - $$x_{ij} := \chi_{ij}\xi_{ij} \tag{2}$$
  - χ_{ij} are iid indicator random variables with $\mathbb{P}(\chi_{ij} = 1) = \theta$
  - ξ_{ij} are iid, mean 0, variance bounded by 1
  - $$\mathbb{P}(|\xi| > t) < 2\exp(-t^2/2).$$
- This model includes many important distributions such as the standard Gaussians and Rademachers.

Uniqueness

- If $Y = AX$, then $Y = (AV)(V^{-1}X)$ for any diagonal matrix V with non-zero diagonal entries.
- One can freely permute the columns of A and the rows of X accordingly while keeping Y the same.
- In the rest of this talk, unique recovery will be understood modulo these two operations.

Uniqueness Result

Theorem 1.

There is a constant $C > 0, \alpha > 0$ such that the following holds. Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $1/n \leq \theta \leq \alpha/\sqrt{n}$ then for $p \geq Cn \log n$ the equation $Y = AX$ has a unique solution, with high probability.

This theorem is essentially Theorem 3 from Spielman, et al., with a slight strengthening that allows us to drop the assumption that the random variables ξ_{ij} have a symmetric distribution.

Regarding efficient recovery (Problem 2.), Spielman et al showed

Theorem 2.

There is a constant $C > 0, \alpha > 0$ such that the following holds. Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $2/n \leq \theta \leq \alpha/\sqrt{n}$. Then for $p \geq Cn^2 \log^2 n$, one can efficiently find a solution with probability $1 - o(1)$.

Quadratic Gap

There is a quadratic gap between the bounds (on p) in Theorems 1 and 2. [4] conjectured that the lower bound is the truth and posed the following problem

Conjecture 3.

There is a constant $C > 0, \alpha > 0$ such that the following holds. Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $2/n \leq \theta \leq \alpha/\sqrt{n}$. Then for $p \geq Cn \log n$, one can efficiently find a solution with probability $1 - o(1)$.

Our main result confirms Conjecture 3 up to a logarithmic term.

Theorem 4 (LV '15).

Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $2/n \leq \theta \leq \alpha/\sqrt{n}$. Then for $p \geq Cn \log^3 n$, one can efficiently find a solution with probability $1 - o(1)$

We achieve this goal by giving a sharper analysis of the algorithm used in Spielman et al. The key tool in our analysis is a new ϵ-net argument, which provides a more effective way to use the union bound and is of independent interest.

* In [1], they show that this result requires a modified version of ER-SpUD.

Refinements

Optimal bound for a regime.

Theorem 5.

Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $c_1/n \leq \theta \leq c_2/n$. Then for $p \geq Cn \log n$, one can efficiently find a solution with probability $1 - o(1)$

Combining the proof of Theorem 4 with a result from random matrix theory, we obtain the following more general result, which handles the case when A is rectangular.

Theorem 6 (LV '15).

Let A be an $n \times m$ matrix of rank m and and X a sparse random $m \times p$ matrix with $2/n \leq \theta \leq \alpha/\sqrt{n}$. Then for $p \geq Cn \log^3 n$, one can efficiently find a solution with probability $1 - o(1)$

Intuition

- Instead of recovering A directly, recover rows of X.
- X and Y have same rowspace.
- The sparsest rows in the rowspace of Y are likely to be rows of X.

ER-SpUD v2 [1]

Algorithm 1 ER-SpUD(DC)v2

1: Create all $T = \binom{p}{2}$ pairings of columns of Y and for $j \in [T]$ let $g_j = \{Ye_{j_1}, Ye_{j_2}\}$
2: For $j = 1, \ldots, T$
Let $r_j = Ye_{j_1} + Ye_{j_2}$, where $g_j = \{Ye_{j_1}, Ye_{j_2}\}$
Solve $min_w \|w^T Y\|_1$ subject to $(Y r_j)^T w = 1$, and set $s_j = w^T Y$.
3: Use Greedy algorithm to reconstruct X and A.

Greedy Algorithm

Algorithm 2 Greedy

1: Require: $S = \{s_1, \ldots, s_T\} \subset \mathbb{R}^p$
2: For $i = 1 \ldots n$
  REPEAT
    $l \leftarrow arg\ min_{s_l \in S} \|s_l\|_0$, breaking ties arbitrarily
    $x_i = s_l$
    $S = S \backslash \{s_l\}$
  UNTIL $rank([x_1, \ldots, x_i]) = i$
3: Set $X = [x_1, \ldots, x_i]^T$, and $A = YY^T(XY^T)^{-1}$

From Dictionary Learning to Matrix Concentration

- By a change of variables $z = A^T w$, $b = A^{-1} r$, we can consider the equivalent problem

$$\text{minimize } \|z^T X\|_1 \text{ subject to } b^T z = 1. \tag{3}$$

- The optimization recovers a row of X if z_* is 1-sparse.
- We prove the sparsity of z_* in stages.

From Dictionary Learning to Matrix Concentration

- By a change of variables $z = A^T w$, $b = A^{-1} r$, we can consider the equivalent problem

$$\text{minimize } \|z^T X\|_1 \text{ subject to } b^T z = 1. \tag{3}$$

- The optimization recovers a row of X if z_* is 1-sparse.
- We prove the sparsity of z_* in stages.

From Dictionary Learning to Matrix Concentration

- By a change of variables $z = A^T w$, $b = A^{-1} r$, we can consider the equivalent problem

$$\text{minimize } \|z^T X\|_1 \text{ subject to } b^T z = 1. \tag{3}$$

- The optimization recovers a row of X if z_* is 1-sparse.
- We prove the sparsity of z_* in stages.

From Dictionary Learning to Matrix Concentration

- In the first stage we show that z_* is supported on the s non-zero entries of b indexed by J .
- Let $z_0 = P_J z_*$ and $z_1 = z_* - z_0$ and S be the indices of the nonzero columns in X_J
- We show this with contradiction:

$$\|z_*^T X\|_1 \geq \|z_0^T X\|_1 - 2\|z_1^T X^S\|_1 + \|z_1^T X\|_1.$$

- Thus, if $\|z_1^T X\|_1 - 2\|z_1^T X^S\|_1 > 0$, then z_0 has a lower objective value.
- Notice that

$$\mathbb{E}[\|z^T X\|_1 - 2\|z^T X^S\|_1] = (p - 2|S|)\mathbb{E}\|z^T X\|_1$$

From Dictionary Learning to Matrix Concentration

- In the first stage we show that z_* is supported on the s non-zero entries of b indexed by J .
- Let $z_0 = P_J z_*$ and $z_1 = z_* - z_0$ and S be the indices of the nonzero columns in X_J
- We show this with contradiction:

$$\|z_*^T X\|_1 \geq \|z_0^T X\|_1 - 2\|z_1^T X^S\|_1 + \|z_1^T X\|_1.$$

- Thus, if $\|z_1^T X\|_1 - 2\|z_1^T X^S\|_1 > 0$, then z_0 has a lower objective value.
- Notice that

$$\mathbb{E}[\|z^T X\|_1 - 2\|z^T X^S\|_1] = (p - 2|S|)\mathbb{E}\|z^T X\|_1$$

From Dictionary Learning to Matrix Concentration

- In the first stage we show that z_* is supported on the s non-zero entries of b indexed by J .
- Let $z_0 = P_J z_*$ and $z_1 = z_* - z_0$ and S be the indices of the nonzero columns in X_J
- We show this with contradiction:

$$\|z_*^T X\|_1 \geq \|z_0^T X\|_1 - 2\|z_1^T X^S\|_1 + \|z_1^T X\|_1.$$

- Thus, if $\|z_1^T X\|_1 - 2\|z_1^T X^S\|_1 > 0$, then z_0 has a lower objective value.
- Notice that

$$\mathbb{E}[\|z^T X\|_1 - 2\|z^T X^S\|_1] = (p - 2|S|)\mathbb{E}\|z^T X\|_1$$

From Dictionary Learning to Matrix Concentration

- In the first stage we show that z_* is supported on the s non-zero entries of b indexed by J .
- Let $z_0 = P_J z_*$ and $z_1 = z_* - z_0$ and S be the indices of the nonzero columns in X_J
- We show this with contradiction:

$$\|z_*^T X\|_1 \geq \|z_0^T X\|_1 - 2\|z_1^T X^S\|_1 + \|z_1^T X\|_1.$$

- Thus, if $\|z_1^T X\|_1 - 2\|z_1^T X^S\|_1 > 0$, then z_0 has a lower objective value.
- Notice that

$$\mathbb{E}[\|z^T X\|_1 - 2\|z^T X^S\|_1] = (p - 2|S|)\mathbb{E}\|z^T X\|_1$$

From Dictionary Learning to Matrix Concentration

- In the first stage we show that z_* is supported on the s non-zero entries of b indexed by J .
- Let $z_0 = P_J z_*$ and $z_1 = z_* - z_0$ and S be the indices of the nonzero columns in X_J
- We show this with contradiction:

$$\|z_*^T X\|_1 \geq \|z_0^T X\|_1 - 2\|z_1^T X^S\|_1 + \|z_1^T X\|_1.$$

- Thus, if $\|z_1^T X\|_1 - 2\|z_1^T X^S\|_1 > 0$, then z_0 has a lower objective value.
- Notice that

$$\mathbb{E}[\|z^T X\|_1 - 2\|z^T X^S\|_1] = (p - 2|S|)\mathbb{E}\|z^T X\|_1$$

Heart of the matter:

$\|X^T v\|_1$ is concentrated around $\mathbb{E}\|X^T v\|_1 := \mu_v$

$$1.1\mu_v \geq \|X^T v\|_1 \geq .9\mu_v,$$

simultaneously for all $v \in \mathbb{R}^n$.

Heart of the matter:

$\|X^T v\|_1$ is concentrated around $\mathbb{E}\|X^T v\|_1 := \mu_v$

$$1.1\mu_v \geq \|X^T v\|_1 \geq .9\mu_v,$$

simultaneously for all $v \in \mathbb{R}^n$.

p needs to be sufficiently large to guarantee concentration.

Heart of the matter:

$\|X^T v\|_1$ is concentrated around $\mathbb{E}\|X^T v\|_1 := \mu_v$

$$1.1\mu_v \geq \|X^T v\|_1 \geq .9\mu_v,$$

simultaneously for all $v \in \mathbb{R}^n$.

p needs to be sufficiently large to guarantee concentration.

X^T is $p \times n$; $v \in \mathbb{R}^n$; $\|X^T v\|_1 = \sum_{i=1}^{p} |X_i \cdot v|$.

The standard ϵ-net argument

$$\mu_{min} = \min_{\|v\|_1=1} \mu_v = p\sqrt{\theta/n}.$$

Let $Bad(v)$ be the event that $|\|X^T v\|_1 - \mu_v\| \geq c\mu_{min}$ $(c = .1)$.

The standard ϵ-net argument

$$\mu_{min} = \min_{\|v\|_1=1} \mu_v = p\sqrt{\theta/n}.$$

Let $Bad(v)$ be the event that $|\|X^T v\|_1 - \mu_v\| \geq c\mu_{min}$ $(c = .1)$.

Goal. If p is sufficiently large, then

$$\mathbb{P}(\cup_{v\in\mathbb{R}^n} Bad(v)) = o(1). \tag{4}$$

The standard ϵ-net argument

$$\mu_{min} = \min_{\|v\|_1=1} \mu_v = p\sqrt{\theta/n}.$$

Let $Bad(v)$ be the event that $|\|X^T v\|_1 - \mu_v\| \geq c\mu_{min}$ $(c = .1)$.

Goal. If p is sufficiently large, then

$$\mathbb{P}(\cup_{v\in\mathbb{R}^n} Bad(v)) = o(1). \tag{4}$$

Assume $x_{ij} = \chi_{ij}\xi_{ij}$, where ξ_{ij} are Rademacher (± 1).

Definition 7.

A set $\mathcal{N} \subset \mathbb{R}^n$ is an ϵ-net of a set $D \subset \mathbb{R}^n$ in l_q norm, for some $0 < q \leq \infty$, if for any $x \in D$ there is $y \in \mathcal{N}$ so that $\|x - y\|_q \leq \epsilon$. The unit sphere in l_q norm consists of vectors v where $\|v\|_q = 1$. B denotes the unit sphere in l_1 norm.

It suffices to consider the vectors in B.

Definition 7.

A set $\mathcal{N} \subset \mathbb{R}^n$ is an ϵ-net of a set $D \subset \mathbb{R}^n$ in l_q norm, for some $0 < q \leq \infty$, if for any $x \in D$ there is $y \in \mathcal{N}$ so that $\|x - y\|_q \leq \epsilon$. The unit sphere in l_q norm consists of vectors v where $\|v\|_q = 1$. B denotes the unit sphere in l_1 norm.

It suffices to consider the vectors in B.

It is easy to show that for any $v \in B$

$$\mu_{min} := p\sqrt{\theta/n} \leq \mu_v \leq p\theta := \mu_{\max},$$

where the lower bound is achieved at $v = \frac{1}{n}\mathbf{1}$ ($\mathbf{1}$ is the all one vector) and the upper bound at $v = (1, 0, \ldots, 0)$.

Let $\mathcal{N}_0$ be the set of all vectors in B whose coordinates are integer multiples of n^{-3}. $\mathcal{N}_0$ is an n^{-2}-net of B in l_1 norm;
$|\mathcal{N}_0| \leq \exp(4n \log n)$.

Let $\mathcal{N}_0$ be the set of all vectors in B whose coordinates are integer multiples of n^{-3}. $\mathcal{N}_0$ is an n^{-2}-net of B in l_1 norm; $|\mathcal{N}_0| \leq \exp(4n \log n)$.

Suffices to show

$$\mathbb{P}(\cup_{v \in \mathcal{N}_0} Bad(v)) = o(1). \tag{5}$$

Bernstein's inequality

Bound $\mathbb{P}(Bad(v))$ for a fix v.

$$\|X^T v\|_1 = \sum_{i=1}^{p} |X_i v|,$$

where X_i are the columns of X.

Lemma 8 (Berstein).

Let $Z_1, \ldots, Z_n$ be independent random variables such that $|Z_i| \leq \tau$ with probability 1. Let $S := \sum_{i=1}^{n} Z_i$. Then for any $T > 0$

$$\max\{\mathbb{P}(|S - \mathbb{E}S| \geq T) \leq 2\exp(-\frac{T^2}{2(\mathbf{Var}S + T\tau)}).$$

$$Z_i = |X_i v| = \sum_{i=1}^{n} x_{ij} v_j.$$

$|x_{ij} = \chi_{ij}\xi_{ij}| \leq 1$. Thus

$$|Z_i| \leq \sum_{i=1}^{n} |v_j| = \|v\|_1 = 1; \;\; \tau = 1.$$

$$\mathbf{Var} \sum_{i=1}^{p} Z_i = p\mathbf{Var} Z_i \leq p\mathbb{E}|X_i v|^2 = p \sum_{j=1}^{n} \theta v_j^2 \leq p\theta \sum_{j=1}^{n} |v_j| = p\theta.$$

Set $T = c\mu_{min} = cp\sqrt{\theta/n}$. Berstein implies

$$\mathbb{P}(Bad(v) \leq 2\exp(-\min\{\frac{c^2 p^2 \theta/n}{4p\theta}, \frac{cp\sqrt{\theta/n}}{4}\}) = 2\exp(-\frac{c^2 p}{4n})$$

Union bound

Using the union bound

$$\mathbb{P}(\cup_{v \in \mathcal{N}_0} Bad(v)) \leq \sum_{v \in \mathcal{N}_0} \mathbb{P}(Bad(v)) \tag{6}$$

we obtain

$$\mathbb{P}(\cup_{v \in \mathcal{N}_0} Bad(v)) \leq |\mathcal{N}_0| \times 2 \exp(-\frac{c^2 p}{4n}).$$

But $\mathcal{N}_0 \leq \exp(4n \log n))$. So $p \geq Cn^2 \log n$ suffices.

A better way to use union bound

Assume that one can split the net $\mathcal{N}_0$ into m disjoint clusters $\mathcal{C}_i$, $1 \leq i \leq m$, so that if u and v belong to the same cluster $\mathbb{P}(Bad(u) \backslash Bad(v)) \leq p_1$, where p_1 is much smaller than p_0, then

$$\mathbb{P}(\cup_{v \in \mathcal{C}_i} Bad(v)) \leq \mathbb{P}(Bad(v^{[i]})) + |\mathcal{C}_i| p_1,$$

where $v^{[i]}$ is a representative point in $\mathcal{C}_i$. Summing over i,

$$\mathbb{P}(\cup_{v \in \mathcal{N}_0} Bad(v)) \leq \sum_{i=1}^{m} \mathbb{P}(Bad(v^{[i]})) + |\mathcal{N}_0| p_1 \leq m p_0 + |\mathcal{N}_0| p_1. \quad (7)$$

A better way to use union bound

Assume that one can split the net $\mathcal{N}_0$ into m disjoint clusters $\mathcal{C}_i$, $1 \leq i \leq m$, so that if u and v belong to the same cluster $\mathbb{P}(Bad(u) \backslash Bad(v)) \leq p_1$, where p_1 is much smaller than p_0, then

$$\mathbb{P}(\cup_{v \in \mathcal{C}_i} Bad(v)) \leq \mathbb{P}(Bad(v^{[i]})) + |\mathcal{C}_i| p_1,$$

where $v^{[i]}$ is a representative point in $\mathcal{C}_i$. Summing over i,

$$\mathbb{P}(\cup_{v \in \mathcal{N}_0} Bad(v)) \leq \sum_{i=1}^{m} \mathbb{P}(Bad(v^{[i]})) + |\mathcal{N}_0| p_1 \leq m p_0 + |\mathcal{N}_0| p_1. \quad (7)$$

We gain significantly if $p_1 \ll p_0$ and $m \ll |\mathcal{N}_0|$.

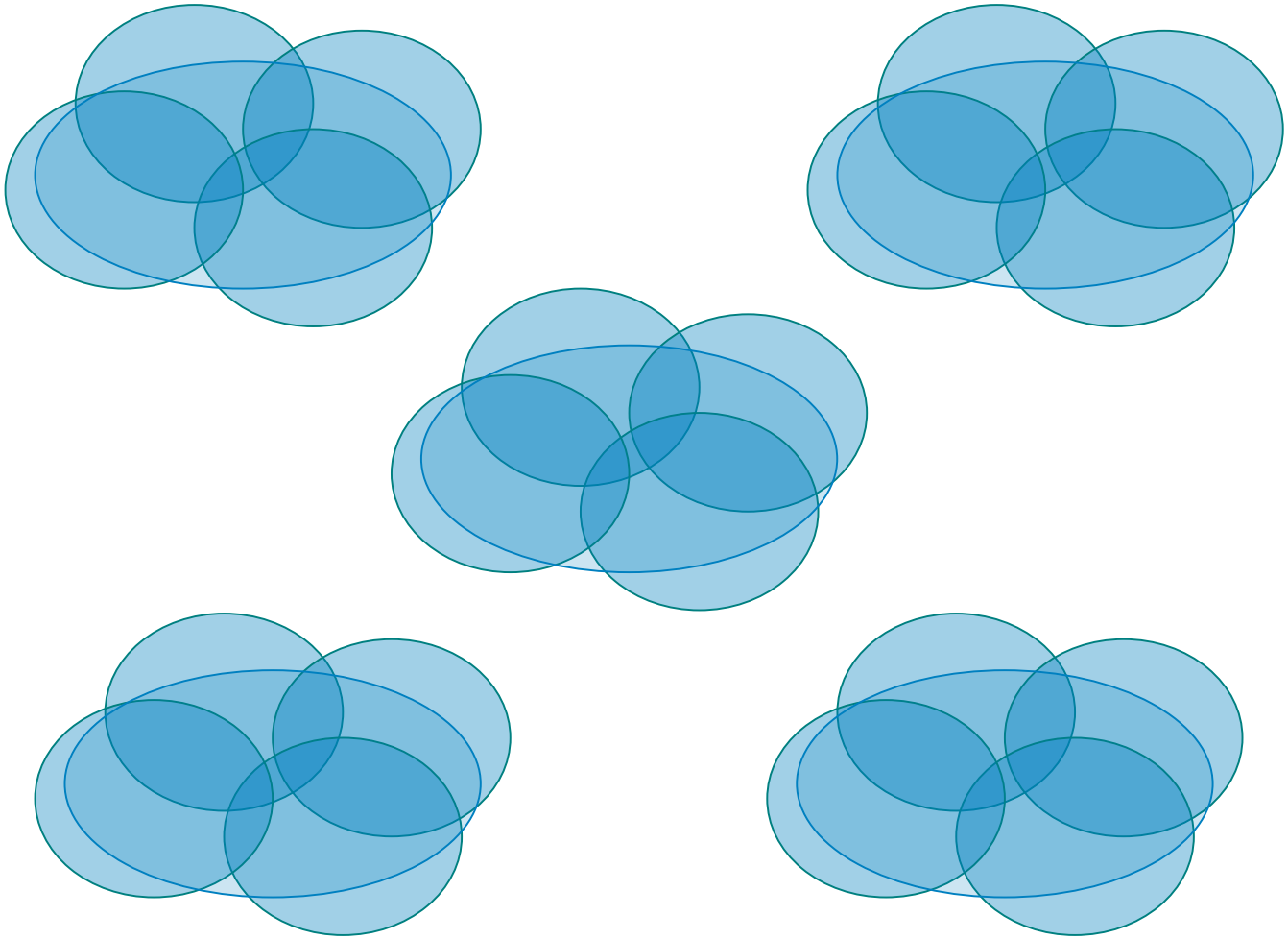


Figure: Bad events appear in clusters

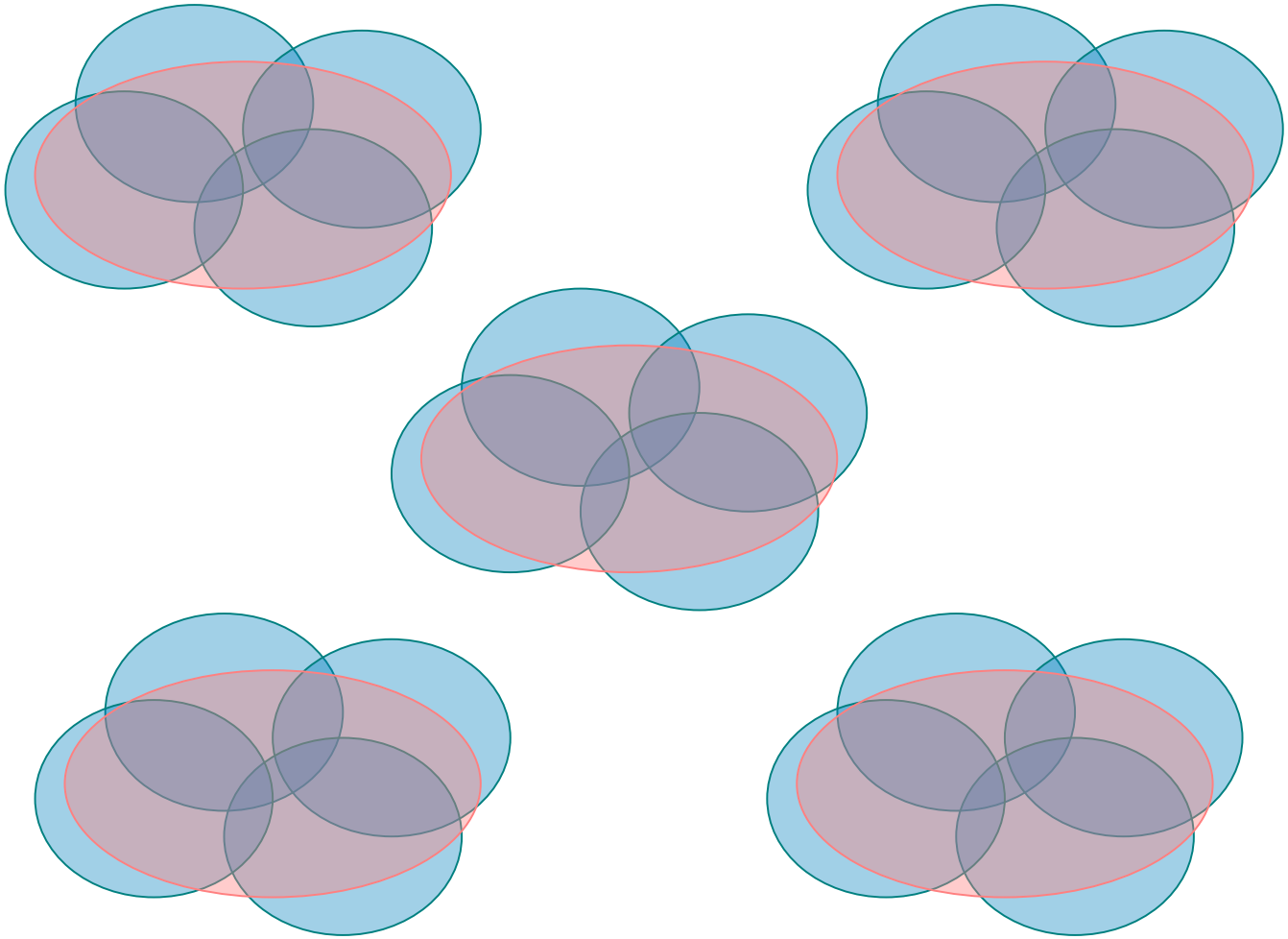


Figure: Choose representative points

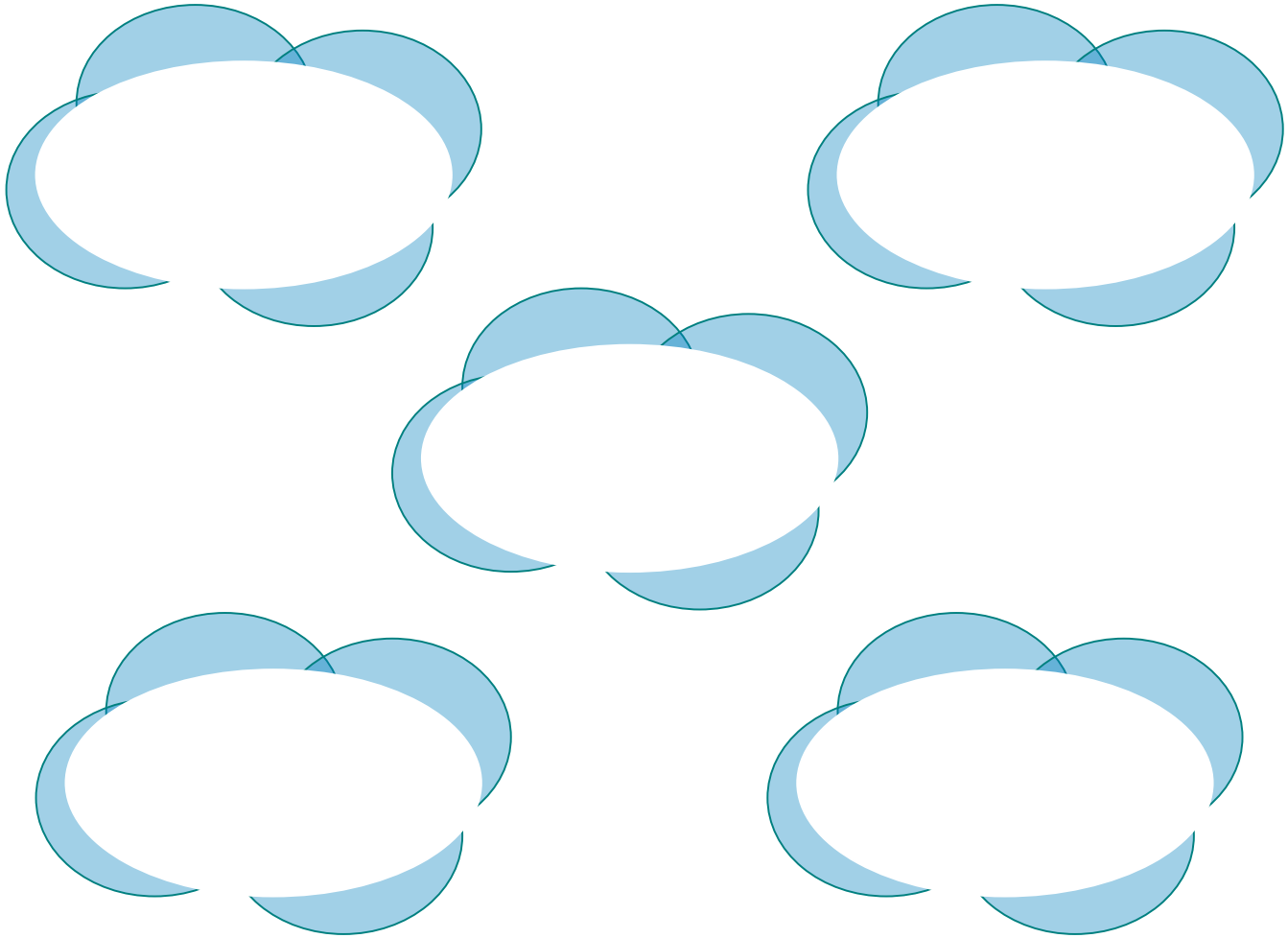


Figure: $\mathbb{P}(Bad(u) \backslash Bad(v)) \leq p_1$, where p_1 is much smaller than p_0

Lemma 9.

Let $\mathcal{P}$ be a probability space. Let $\mathcal{N} = \mathcal{N}_0$ be a finite set, where to each element $v \in \mathcal{N}_0$ we associate a set $Bad_0(v) \subset \mathcal{P}$. Assume that we can construct a sequence of sets

$$\mathcal{N}_L, \mathcal{N}_{L-1}, \ldots, \mathcal{N}_0,$$

and for each $u \in \mathcal{N}_l, 1 \leq l \leq L$ an event $Bad_l(u)$ such that the following holds. For each $v \in \mathcal{N}_{l-1}$, there is $u \in \mathcal{N}_l$ such that $\mathbb{P}(Bad_{l-1}(v) \backslash Bad_l(u)) \leq p_l$ and for each $u \in \mathcal{N}_L$, $\mathbb{P}(Bad_L(u)) \leq p_0$. Then

$$\mathbb{P}(\cup_{v \in \mathcal{N}_0} Bad_0(v)) \leq |\mathcal{N}_L| p_0 + \sum_{l=1}^{L} |\mathcal{N}_{l-1}| p_l.$$

The construction of $\mathcal{N}_l$ are of critical importance, and we are going to construct them using the l_∞ distance, rather than the obvious choice of l_1. *This is the key point of our method.*

Strong concentration

Lemma 10.

If $p \geq n \log^4 n$, then with probability $1 - o(1)$, for all $v \in B$

$$|\|X^T v\|_1 - \mathbb{E}\|X^T v\|_1| \leq c\mu_{min}.$$

Toy argument: union bound with two levels

$\mathcal{N}_0$ consists of vectors with unit l_1 norm and coordinates integer multiples of n^{-3}.

Toy argument: union bound with two levels

$\mathcal{N}_0$ consists of vectors with unit l_1 norm and coordinates integer multiples of n^{-3}.

$\mathcal{N}_1'$ consists of vectors with unit l_1 norm and coordinates integer multiples of $n^{-1/2}$.

Toy argument: union bound with two levels

$\mathcal{N}_0$ consists of vectors with unit l_1 norm and coordinates integer multiples of n^{-3}.

$\mathcal{N}_1'$ consists of vectors with unit l_1 norm and coordinates integer multiples of $n^{-1/2}$.

Use $\mathcal{N}_1'$ to cluster $\mathcal{N}_0$. Use the points in $\mathcal{N}_1'$ as centered to Voronoi partition $\mathcal{N}_0$, with respect to the l_∞ **norm.**
Further partition each Voronoi cell into a few parts so that in each part the expectations μ are close to each other

$$\mu_w - \mu_v \leq \frac{c}{4}\mu_{min}.$$

Toy argument: union bound with two levels

$\mathcal{N}_0$ consists of vectors with unit l_1 norm and coordinates integer multiples of n^{-3}.

$\mathcal{N}_1'$ consists of vectors with unit l_1 norm and coordinates integer multiples of $n^{-1/2}$.

Use $\mathcal{N}_1'$ to cluster $\mathcal{N}_0$. Use the points in $\mathcal{N}_1'$ as centered to Voronoi partition $\mathcal{N}_0$, with respect to the l_∞ **norm.**
Further partition each Voronoi cell into a few parts so that in each part the expectations μ are close to each other

$$\mu_w - \mu_v \leq \frac{c}{4}\mu_{min}.$$

For each set in the final partition, choose a point u. These u form $\mathcal{N}_1$.

Repeat the union bound argument (with $c/2$) for $\mathcal{N}_1$. Since

$$|\mathcal{N}_1| \approx |\mathcal{N}_1'| \leq \binom{n}{n^{1/2}} (n^{1/2})^{n^{1/2}} \leq \exp(Cn^{1/2} \log n),$$

we only need $p = n^{3/2+o(1)}$.

Repeat the union bound argument (with $c/2$) for $\mathcal{N}_1$. Since

$$|\mathcal{N}_1| \approx |\mathcal{N}_1'| \leq \binom{n}{n^{1/2}} (n^{1/2})^{n^{1/2}} \leq \exp(Cn^{1/2} \log n),$$

we only need $p = n^{3/2+o(1)}$.
If for all $u \in \mathcal{N}_1$,

$$|X^T u - \mu_u| \leq \frac{c}{2} \mu_{min}$$

but there is $v \in \mathcal{N}_0$

$$|X^T v - \mu_v| \geq c\mu_{min},$$

then by choosing u *close* to v

Repeat the union bound argument (with $c/2$) for $\mathcal{N}_1$. Since

$$|\mathcal{N}_1| \approx |\mathcal{N}_1'| \leq \binom{n}{n^{1/2}} (n^{1/2})^{n^{1/2}} \leq \exp(Cn^{1/2} \log n),$$

we only need $p = n^{3/2+o(1)}$.
If for all $u \in \mathcal{N}_1$,

$$|X^T u - \mu_u| \leq \frac{c}{2}\mu_{min}$$

but there is $v \in \mathcal{N}_0$

$$|X^T v - \mu_v| \geq c\mu_{min},$$

then by choosing u *close* to v

$$|X^T v - X^T u| \geq \frac{c}{2}\mu_{min} - |\mu_u - \mu_v| \geq \frac{c}{4}\mu_{min}.$$

$$X^T v - X^T u = X^T(v-u) = \sum_{i=1}^{p} X_i^T(v-u) = \sum_{i=1}^{p} Z_i.$$

$$\mathbf{Var} Z_i \leq \theta \sum_{j=1}^{n} (v_j - u_j)^2 \leq 2\theta \|v-u\|_\infty \leq 2\theta n^{-1/2}.$$

$$\mathbf{Var} \sum_{i=1}^{p} Z_i = p\mathbf{Var} Z_i \leq 2p\theta n^{-1/2}.$$

Set $T = \frac{c}{4}\mu_{min} = \frac{c}{4} p\sqrt{\theta/n}$. Berstein implies

$$\mathbb{P}(Bad(v) \leq 2\exp(-\min\{\frac{c^2 p^2 \theta/n}{64 p n^{-1/2}\theta}, \frac{cp\sqrt{\theta/n}}{64}\}).$$

$$\mathbb{P}(Bad(v) \leq 2\exp(-\min\{\frac{c^2p^2\theta/n}{64pn^{-1/2}\theta}, \frac{cp\sqrt{\theta/n}}{64}\}).$$

In the first term $\frac{c^2p^2\theta/n}{64pn^{-1/2}\theta}$ we gain an extra $n^{1/2}$ as the variance gets smaller; so $p = n^{3/2+o(1)}$ suffices.

To deal with the second term one needs to refine the Berstein bound itself.

Undercomplete Dictionaries

There are many situations where the data has high ambient dimension, but lives in a low-dimensional subspace.
Such situations occur in

- Signal Processing
- Feature Recognition
- Machine Learning

Rectangular Dictionaries

Consider a full rank matrix A of size $n > m$, such that $n - m = \omega(1)$, and the equation $AX = Y$.

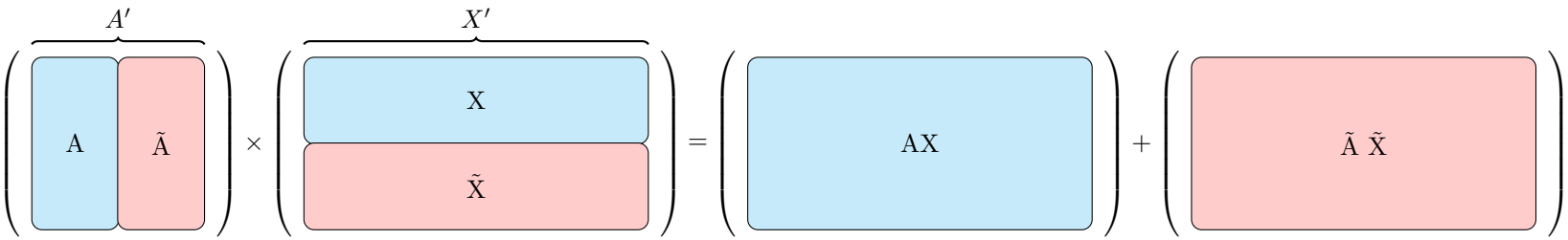


Figure: Rectangular A with $n > m$

A random matrix result

Theorem 11 (Bourgain-V.-Wood 2008).

For every $\epsilon > 0$ there exists $\delta > 0$ such that the following holds. Let $N_{f,n}$ be an n by n complex matrix in which f rows contain fixed, non-random entries and where the other rows contain entries that are independent discrete random variables. If the fixed rows have co-rank k and if for every random entry α, we have $max_x \mathbb{P}(\alpha = x) \leq 1 - \epsilon$, then for all sufficiently large n

$$\mathbb{P}(N_{f,n} \text{ has co-rank} > k) \leq (1 - \delta)^{n-f}.$$

Recent Work

1. Adamczak, Radosaw. "A note on the sample complexity of the Er-SpUD algorithm by Spielman, Wang and Wright for exact recovery of sparsely used dictionaries." arXiv preprint arXiv:1601.02049 (2016).
2. Basiok, Jarosaw, and Jelani Nelson. "An improved analysis of the ER-SpUD dictionary learning algorithm." arXiv preprint arXiv:1602.05719 (2016).

Using powerful chaining arguments, both works removed the extra $\log^3 n$ factor in the sample complexity.

Main Improvement [1]

Theorem 12 (Blasiok, Nelson '16).

When $p = \Omega(\varepsilon^{-2} n \log \frac{n}{\delta})$,

$$\mathbb{P}\left(\sup_{v \in B_1} \left| \|X^T v\|_1 - \mathbb{E}\|X^T v\|_1 \right| > \varepsilon \mathbb{E}\|X^T v\|_1 \right) < \delta \qquad (8)$$

Admissible Sequences and Gamma Functionals

Definition 13 (Admissible sequence).

For an arbitrary set T, we say that a sequence of its subsets $(T_k)_{k=0}^{\infty}$ is admissible if for every number k it is true that $T_k \subset T_{k+1}$ and $|T_k| \leq 2^{2^k}$ for $k \geq 1$ and $|T_0| = 1$.

Definition 14 (Gamma functionals).

For a metric space (T, d) we define

$$\gamma_\alpha(T, d) := \inf_{(T_k)} \sup_{x \in T} \sum_{k=0}^{\infty} 2^{k/\alpha} d(x, T_k) \tag{9}$$

where the infimum is taken over all admissible sequences T_k. In the above formula we define as usual $d(x, T_k) := \inf_{t \in T_k} d(x, t)$.

Some Facts

Theorem 15 (Majorizing measures, Talagrand '14).

Let $T \subset \mathbb{R}^n$, and assume that $g = (g_1, \dots g_n)$ is a vector of i.i.d. standard normal random variables. Then

$$\mathbb{E} \sup_{t \in T} \langle g, t \rangle \simeq \gamma_2(T, d_2) \tag{10}$$

Where d_p is the metric induced by the ℓ_p norm.

Theorem 16 (Talagrand '14, Theorem 10.2.8).

Let $T \subset \mathbb{R}^n$, and assume that $x = (x_1, \dots, x_n)$ is a vector of i.i.d. standard exponential random variables. Then

$$\mathbb{E} \sup_{t \in T} \langle t, x \rangle \simeq \gamma_2(T, d_2) + \gamma_1(T, d_\infty) \tag{11}$$

Theorem 17 (Generic chaining (Talagrand), Theorem 2.2.23).

Let T be an arbitrary set of indices, and $d_1, d_2 : T \times T \to \mathbb{R}_{\geq 0}$ two metrics on T. Suppose that with any point $t \in T$ we have associated random variable X_t, with $\mathbb{E}X_t = 0$. Suppose moreover, that for any two points $u, w \in T$ we have a tail bound:

$$\mathbb{P}\left(|X_u - X_v| > \lambda\right) \lesssim \exp\left(-\frac{\lambda^2}{d_1(u,v)^2}\right) + \exp\left(-\frac{\lambda}{d_2(u,v)}\right)$$

Then

$$\mathbb{E}\sup_{u \in T} |X_u| \lesssim \gamma_2(T, d_1) + \gamma_1(T, d_2)$$

Theorem 18 (Dirksen, '15).

Let T be an arbitrary set of indices and $d_1, d_2 : T \times T \to \mathbb{R}_{\geq 0}$ two metrics on T. Suppose that with any point $t \in T$ we have associated random variable X_t, such that $\mathbb{E}X_t = 0$. Suppose moreover that for any two points $u, w \in T$, we have a tail bound

$$\mathbb{P}(|X_u - X_v| > \lambda) \lesssim \exp\left(-\frac{\lambda^2}{d_1(u,v)^2}\right) + \exp\left(-\frac{\lambda}{d_2(u,v)}\right)$$

Then for there exists a constant C, such that for any $u > 0$

$$\mathbb{P}(\sup_{u \in T} |X_u| > C(\gamma_2(T, d_1) + \gamma_1(T, d_2) + \sqrt{u}\Delta(T, d_1) + u\Delta(T, d_2)) < e^{-u}$$

where $\Delta(T, d) := \sup_{u,v \in T} d(u, v)$.

Fact 19.

$\gamma_2(B_1, d_2) \lesssim \sqrt{\log n}$

Proof.

By Theorem 15, we have

$$\gamma_2(B_1, d_2) \lesssim \mathbb{E}_g \sup_{t \in B_1} \langle t, g \rangle \tag{12}$$

where g is a Gaussian vector. By the duality of ℓ_1 and ℓ_∞ norms, for any vector $w \in \mathbb{R}^n$ we have $\sup_{t \in B_1} \langle t, v \rangle = \|v\|_\infty$, so in particular $\mathbb{E} \sup_{t \in B_1} \langle t, g \rangle = \mathbb{E}\|g\|_\infty \simeq \sqrt{\log n}$. □

Fact 20.

$\gamma_1(B_1, d_\infty) \lesssim \log n$

One can show that for every pair of points $u, v \in B_1$, we have

$$\mathbb{P}(|\|X^T u\|_1 - \|X^T v\|_1| > \lambda) \lesssim \exp\left(-\frac{\lambda^2}{2m\theta\|u-v\|_2^2}\right) + \exp\left(-\frac{\lambda}{\|u-v\|_\infty}\right)$$

The proof follows from a modification of Bernstein's inequality or properties of subgamma random variables.

Combining the previous slide, the two facts, and Dirksen's tail bound gives

$$\mathbb{P}\left(sup_{v\in B_1}|\|X^T v\|_1 - \mathbb{E}\|X^T v\|_1| < L_2\left(\sqrt{p\theta(\log n + \log\frac{1}{\delta})} + \log n + \log\frac{1}{\delta}\right)\right) < \delta$$

where L_2 is some constant. A simple calculation shows that $p > C\varepsilon^2 n\log(n/\delta)$ suffices to make the probability less than δ.

Acknowledgements

We would like to thank Dan Spielman for bringing this problem to our attention.

 Basiok, Jarosaw, and Jelani Nelson. An improved analysis of the ER-SpUD dictionary learning algorithm. *arXiv preprint arXiv:1602.05719* (2016).

 Mairal, Julien, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *arXiv preprint arXiv:1411.3230* (2014).

 Adamczak, Radosaw. A note on the sample complexity of the Er-SpUD algorithm by Spielman, Wang and Wright for exact recovery of sparsely used dictionaries. *arXiv preprint arXiv:1601.02049* (2016).

 Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. *arXiv preprint arXiv:1206.5882*, 2012.

Thank you!