# **Decoding in Compressed Sensing**

Albert Cohen, Wolfgang Dahmen, Ronald DeVore, Guergana Petrova, Przemek Wojtaszczek



•  $x \in \mathbb{R}^N$  with N large



- $x \in \mathbb{R}^N$  with N large
- We are able to ask n questions about x



- $x \in \mathbb{R}^N$  with N large
- We are able to ask n questions about x
- Question means inner product  $v \cdot x$  with  $v \in \mathbb{R}^N$  called sample

- $x \in \mathbb{R}^N$  with N large
- We are able to ask n questions about x
- Question means inner product  $v \cdot x$  with  $v \in \mathbb{R}^N$  called sample
- What are the best questions to ask??

- $x \in \mathbb{R}^N$  with N large
- We are able to ask n questions about x
- Question means inner product  $v \cdot x$  with  $v \in \mathbb{R}^N$  called sample
- What are the best questions to ask??
- Any such sampling is given by  $\Phi x$  where  $\Phi$  is an  $n \times N$  matrix

- $x \in \mathbb{R}^N$  with N large
- We are able to ask n questions about x
- Question means inner product  $v \cdot x$  with  $v \in \mathbb{R}^N$  called sample
- What are the best questions to ask??
- Any such sampling is given by  $\Phi x$  where  $\Phi$  is an  $n \times N$  matrix
- We are interested in the good / best matrices  $\Phi$

- $x \in \mathbb{R}^N$  with N large
- We are able to ask n questions about x
- Question means inner product  $v \cdot x$  with  $v \in \mathbb{R}^N$  called sample
- What are the best questions to ask??
- Any such sampling is given by  $\Phi x$  where  $\Phi$  is an  $n \times N$  matrix
- We are interested in the good / best matrices  $\Phi$
- Here good means the samples  $y = \Phi x$  contain enough information to approximate x well

• We view  $\Phi$  as an encoder



- We view  $\Phi$  as an encoder
- Since  $\Phi: \mathbb{R}^N \to \mathbb{R}^n$  many x are encoded with same y

TAMU 2008 – p  $3/2^{2}$ 

- We view  $\Phi$  as an encoder
- Since  $\Phi: \mathbb{R}^N \to \mathbb{R}^n$  many x are encoded with same y

TAMU 2008 – p  $3/2^{\circ}$ 

•  $\mathcal{N} := \{\eta : \Phi \eta = 0\}$  the null space of  $\Phi$ 

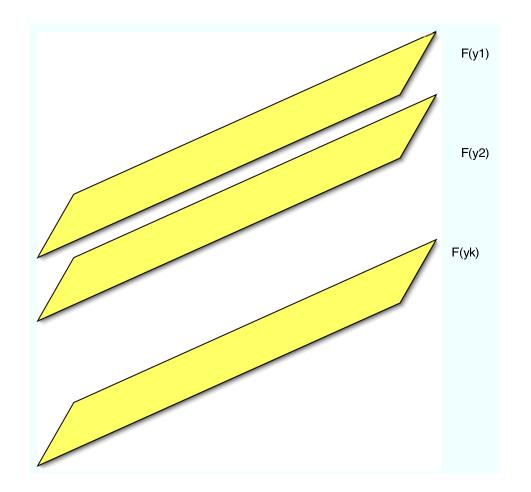
- We view  $\Phi$  as an encoder
- Since  $\Phi: \mathbb{R}^N \to \mathbb{R}^n$  many x are encoded with same y
- $\mathcal{N} := \{\eta : \Phi \eta = 0\}$  the null space of  $\Phi$
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N} \text{ for any } x_0 \in \mathcal{F}(y)$

We view  $\Phi$  as an encoder

• Since  $\Phi: \mathbb{R}^N \to \mathbb{R}^n$  many x are encoded with same y

- $\mathcal{N} := \{\eta : \Phi \eta = 0\}$  the null space of  $\Phi$
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N} \text{ for any } x_0 \in \mathcal{F}(y)$
- The hyperplanes  $\mathcal{F}(y)$  with  $y \in \mathbb{R}^n$  stratify  $\mathbb{R}^N$

The sets  $\mathcal{F}(y)$ 



TAMU 2008 – p. 4/2<sup>2</sup>

We view  $\Phi$  as an encoder

• Since  $\Phi: \mathbb{R}^N \to \mathbb{R}^n$  many x are encoded with same y

- $\mathcal{N} := \{\eta : \Phi \eta = 0\}$  the null space of  $\Phi$
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N} \text{ for any } x_0 \in \mathcal{F}(y)$
- The hyperplanes  $\mathcal{F}(y)$  with  $y \in \mathbb{R}^n$  stratify  $\mathbb{R}^N$
- Decoder is any (possibly nonlinear) mapping  $\Delta$  from  $I\!\!R^n \to I\!\!R^N$

We view  $\Phi$  as an encoder

• Since  $\Phi: \mathbb{R}^N \to \mathbb{R}^n$  many x are encoded with same y

- $\mathcal{N} := \{\eta : \Phi \eta = 0\}$  the null space of  $\Phi$
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N} \text{ for any } x_0 \in \mathcal{F}(y)$
- The hyperplanes  $\mathcal{F}(y)$  with  $y \in \mathbb{R}^n$  stratify  $\mathbb{R}^N$
- Decoder is any (possibly nonlinear) mapping  $\Delta$  from  $I\!\!R^n \to I\!\!R^N$

TAMU 2008 – p. 5/2<sup>•</sup>

•  $\bar{x} := \Delta(\Phi(x))$  is our approximation to x from the information extracted

We view  $\Phi$  as an encoder

• Since  $\Phi: \mathbb{R}^N \to \mathbb{R}^n$  many x are encoded with same y

- $\mathcal{N} := \{\eta : \Phi \eta = 0\}$  the null space of  $\Phi$
- $\mathcal{F}(y) := \{x : \Phi x = y\} = x_0 + \mathcal{N} \text{ for any } x_0 \in \mathcal{F}(y)$
- The hyperplanes  $\mathcal{F}(y)$  with  $y \in \mathbb{R}^n$  stratify  $\mathbb{R}^N$
- Decoder is any (possibly nonlinear) mapping  $\Delta$  from  $I\!\!R^n \to I\!\!R^N$

TAMU 2008 – p. 5/2<sup>•</sup>

- $\bar{x} := \Delta(\Phi(x))$  is our approximation to x from the information extracted
- This is a typical inverse problem since x is underdetermined by y

The only reason such a system might work is because we model signals as sparse or compressible

- The only reason such a system might work is because we model signals as sparse or compressible
- How should we measure performance of the encoding-decoding?

- The only reason such a system might work is because we model signals as sparse or compressible
- How should we measure performance of the encoding-decoding?
- Define  $\Sigma_k := \{x \in \mathbb{R}^N : \# \operatorname{supp}(\mathbf{x}) \le \mathbf{k}\}$

$$\sigma_k(x)_X := \inf_{z \in \Sigma_k} \|x - z\|_X$$

- The only reason such a system might work is because we model signals as sparse or compressible
- How should we measure performance of the encoding-decoding?
- Define  $\Sigma_k := \{x \in \mathbb{R}^N : \# \operatorname{supp}(\mathbf{x}) \le \mathbf{k}\}$

$$\sigma_k(x)_X := \inf_{z \in \Sigma_k} \|x - z\|_X$$

• Given an encoding - decoding pair  $(\Phi, \Delta)$ , we say that this pair is Instance-Optimal of order k for X if for an absolute constant C > 0

TAMU 2008 – p. 6/2<sup>,</sup>

 $||x - \Delta(\Phi(x))||_X \le C\sigma_k(x)_X$ 

- The only reason such a system might work is because we model signals as sparse or compressible
- How should we measure performance of the encoding-decoding?
- Define  $\Sigma_k := \{x \in \mathbb{R}^N : \# \operatorname{supp}(x) \le k\}$

$$\sigma_k(x)_X := \inf_{z \in \Sigma_k} \|x - z\|_X$$

• Given an encoding - decoding pair  $(\Phi, \Delta)$ , we say that this pair is Instance-Optimal of order k for X if for an absolute constant C > 0

 $||x - \Delta(\Phi(x))||_X \le C\sigma_k(x)_X$ 

Given n, N, the best encoding - decoding pairs are those which have the largest k.

TAMU 2008 – p  $6/2^{2}$ 

• What properties determine whether  $(\Phi, \Delta)$  is good?

- What properties determine whether  $(\Phi, \Delta)$  is good?
- One property which is known to be sufficient in other results is the following Restricted Isometry Property (RIP) introduced by Candes-Romberg-Tao

TAMU 2008 – p. 7/2′

- What properties determine whether  $(\Phi, \Delta)$  is good?
- One property which is known to be sufficient in other results is the following Restricted Isometry Property (RIP) introduced by Candes-Romberg-Tao
- ▶  $\Phi$  has the RIP of order *k* with constant  $\delta \in (0, 1)$  if

 $(1-\delta)\|x\|_{\ell_2} \le \|\Phi x\|_{\ell_2} \le (1+\delta)\|x\|_{\ell_2} \quad x \in \Sigma_k$ 

- What properties determine whether  $(\Phi, \Delta)$  is good?
- One property which is known to be sufficient in other results is the following Restricted Isometry Property (RIP) introduced by Candes-Romberg-Tao
- $\Phi$  has the RIP of order k with constant  $\delta \in (0,1)$  if

 $(1-\delta)\|x\|_{\ell_2} \le \|\Phi x\|_{\ell_2} \le (1+\delta)\|x\|_{\ell_2} \quad x \in \Sigma_k$ 

General Philosophy: the larger the value of k the better the performance of  $\Phi$ 

- What properties determine whether  $(\Phi, \Delta)$  is good?
- One property which is known to be sufficient in other results is the following Restricted Isometry Property (RIP) introduced by Candes-Romberg-Tao
- $\Phi$  has the RIP of order *k* with constant  $\delta \in (0, 1)$  if

 $(1-\delta)\|x\|_{\ell_2} \le \|\Phi x\|_{\ell_2} \le (1+\delta)\|x\|_{\ell_2} \quad x \in \Sigma_k$ 

- General Philosophy: the larger the value of k the better the performance of  $\Phi$
- Given n, N, we know there are matrices  $\Phi$  which have RIP for  $k \leq c_0 n / \log(N/n)$

TAMU 2008 – p  $7/2^{\circ}$ 

- What properties determine whether  $(\Phi, \Delta)$  is good?
- One property which is known to be sufficient in other results is the following Restricted Isometry Property (RIP) introduced by Candes-Romberg-Tao
- has the RIP of order k with constant  $\delta \in (0, 1)$  if

 $(1-\delta)\|x\|_{\ell_2} \le \|\Phi x\|_{\ell_2} \le (1+\delta)\|x\|_{\ell_2} \quad x \in \Sigma_k$ 

- General Philosophy: the larger the value of k the better the performance of  $\Phi$
- Given n, N, we know there are matrices  $\Phi$  which have RIP for  $k \leq c_0 n / \log(N/n)$
- This range of k cannot be improved: from now on we refer to this as the largest range of k

• Cohen-Dahmen-DeVore If  $\Phi$  has RIP for 2k then  $\Phi$  is instance-optimal of order k in  $\ell_1$ :

 $||x - \Delta(\Phi(x))||_{\ell_1} \le C\sigma_k(x)_{\ell_1}$ 



• Cohen-Dahmen-DeVore If  $\Phi$  has RIP for 2k then  $\Phi$  is instance-optimal of order k in  $\ell_1$ :

 $||x - \Delta(\Phi(x))||_{\ell_1} \le C\sigma_k(x)_{\ell_1}$ 

This means that there is some decoder - not necessarily practical (we shall discuss practical decoders in a moment)

• Cohen-Dahmen-DeVore If  $\Phi$  has RIP for 2k then  $\Phi$  is instance-optimal of order k in  $\ell_1$ :

 $||x - \Delta(\Phi(x))||_{\ell_1} \le C\sigma_k(x)_{\ell_1}$ 

- This means that there is some decoder not necessarily practical (we shall discuss practical decoders in a moment)
- Hence we know there are matrices  $\Phi$  which have Instance Optimality in  $\ell_1$  for  $k \le c_0 n / \log(N/n)$

• Cohen-Dahmen-DeVore If  $\Phi$  has RIP for 2k then  $\Phi$  is instance-optimal of order k in  $\ell_1$ :

 $||x - \Delta(\Phi(x))||_{\ell_1} \le C\sigma_k(x)_{\ell_1}$ 

TAMU 2008 — p. 8/2′

- This means that there is some decoder not necessarily practical (we shall discuss practical decoders in a moment)
- Hence we know there are matrices  $\Phi$  which have Instance Optimality in  $\ell_1$  for  $k \le c_0 n / \log(N/n)$
- This range of k cannot be improved

• Given n, N and a constant  $C_0$  then we have instance optimality in  $\ell_2$  for k and this  $C_0$  only if  $k \leq \frac{C_0 n}{N}$ 

- Given n, N and a constant  $C_0$  then we have instance optimality in  $\ell_2$  for k and this  $C_0$  only if  $k \leq \frac{C_0 n}{N}$
- This shows that instance-optimal is not a viable concept for  $\ell_2$

TAMU 2008 – p. 9/21

- Given n, N and a constant  $C_0$  then we have instance optimality in  $\ell_2$  for k and this  $C_0$  only if  $k \leq \frac{C_0 n}{N}$
- This shows that instance-optimal is not a viable concept for  $\ell_2$
- We have to make cN measurement to even get instance optimality for k = 1

- Given n, N and a constant  $C_0$  then we have instance optimality in  $\ell_2$  for k and this  $C_0$  only if  $k \leq \frac{C_0 n}{N}$
- This shows that instance-optimal is not a viable concept for  $\ell_2$
- We have to make cN measurement to even get instance optimality for k = 1
- This bound cannot be improved

# **Sample Results:** $\ell_p$

- Given n, N and a constant  $C_0$  then we have instance optimality in  $\ell_2$  for k and this  $C_0$  only if  $k \leq \frac{C_0 n}{N}$
- This shows that instance-optimal is not a viable concept for  $\ell_2$

TAMU 2008 – p. 9/2'

- We have to make cN measurement to even get instance optimality for k = 1
- This bound cannot be improved
- For 1 the range of <math>k is  $k \le c_0 N^{\frac{2-2/p}{1-2/p}} [n/\log(N/n)]^{\frac{p}{2-p}}$

• In other words: Which matrices give biggest range of k?

- In other words: Which matrices give biggest range of k?
- A sufficient condition is that the matrix satisfy RIP or order 2k for k in the large range

- In other words: Which matrices give biggest range of k?
- A sufficient condition is that the matrix satisfy RIP or order 2k for k in the large range
- With high probability on the draw of the matrix any of the following is known to give a best matrix

- In other words: Which matrices give biggest range of k?
- A sufficient condition is that the matrix satisfy RIP or order 2k for k in the large range
- With high probability on the draw of the matrix any of the following is known to give a best matrix
  - Choose at random N vectors from the unit sphere in  $\mathbb{R}^n$  and use these as the columns of  $\Phi$

- In other words: Which matrices give biggest range of k?
- A sufficient condition is that the matrix satisfy RIP or order 2k for k in the large range
- With high probability on the draw of the matrix any of the following is known to give a best matrix
  - Choose at random N vectors from the unit sphere in  $\mathbb{R}^n$  and use these as the columns of  $\Phi$
  - Choose each entry of  $\Phi$  independently and at random from the Gaussian distribution  $\mathcal{N}(0, 1/\sqrt{n})$

- In other words: Which matrices give biggest range of k?
- A sufficient condition is that the matrix satisfy RIP or order 2k for k in the large range
- With high probability on the draw of the matrix any of the following is known to give a best matrix
  - Choose at random N vectors from the unit sphere in  $\mathbb{R}^n$  and use these as the columns of  $\Phi$
  - Choose each entry of  $\Phi$  independently and at random from the Gaussian distribution  $\mathcal{N}(0, 1/\sqrt{n})$
  - We choose each entry of 
     independently and at random from the Bernouli distribution and then normalize columns to have length one.

- In other words: Which matrices give biggest range of k?
- A sufficient condition is that the matrix satisfy RIP or order 2k for k in the large range
- With high probability on the draw of the matrix any of the following is known to give a best matrix
  - Choose at random N vectors from the unit sphere in  $\mathbb{R}^n$  and use these as the columns of  $\Phi$
  - Choose each entry of  $\Phi$  independently and at random from the Gaussian distribution  $\mathcal{N}(0, 1/\sqrt{n})$
  - We choose each entry of 
     independently and at random from the Bernouli distribution and then normalize columns to have length one.
- Problem: There are no known constructions. Moreover verifying RIP is not feasible computationally

TAMU 2008 – p. 10/2

# **Instance-Optimality in Probability**

- We saw that Instance-Optimality for  $\ell_2^N$  is not viable
- We shall now see it is possible if you accept some probability of failure
- Suppose  $\Phi(\omega)$  is a collection of random matrices
- We say this family satisfies RIP of order k with probability  $1 \epsilon$  if a random draw  $\{\Phi(\omega)\}$  will satisfy RIP of order k with probability  $1 \epsilon$
- We say  $\{\Phi(\omega)\}$  is bounded with probability  $1 \epsilon$  if there is an absolute constant  $C_0$  such that for each  $x \in \mathbb{R}^N$

#### $\|\Phi(\omega)(x)\|_{\ell_2^N} \le C_0 \|x\|_{\ell_2^N}$

with probability  $1 - \epsilon$  a random draw  $\{\Phi(\omega)\}$ 

• If  $\{\Phi(\omega)\}$  satisfies RIP of order 3k and boundedness each with probability  $1 - \epsilon$  then there are decoders  $\Delta(\omega)$ such that given any  $x \in \ell_2^N$  we have with probability  $1 - 2\epsilon$ 

 $\|x - \Delta(\omega)\Phi(\omega)(x)\|_{\ell_2^N} \le C_0 \sigma_k(x)_{\ell_2^N}$ 



• If  $\{\Phi(\omega)\}\$  satisfies RIP of order 3k and boundedness each with probability  $1 - \epsilon$  then there are decoders  $\Delta(\omega)$ such that given any  $x \in \ell_2^N$  we have with probability  $1 - 2\epsilon$ 

 $\|x - \Delta(\omega)\Phi(\omega)(x)\|_{\ell_2^N} \le C_0 \sigma_k(x)_{\ell_2^N}$ 

Instance-optimality in probability



• If  $\{\Phi(\omega)\}$  satisfies RIP of order 3k and boundedness each with probability  $1 - \epsilon$  then there are decoders  $\Delta(\omega)$ such that given any  $x \in \ell_2^N$  we have with probability  $1 - 2\epsilon$ 

 $\|x - \Delta(\omega)\Phi(\omega)(x)\|_{\ell_2^N} \le C_0 \sigma_k(x)_{\ell_2^N}$ 

TAMU 2008 – p 12/2

- Instance-optimality in probability
- Theorem holds for Gaussian, Bernouli and other random families of matrices with  $\epsilon = e^{-cn}$

• If  $\{\Phi(\omega)\}$  satisfies RIP of order 3k and boundedness each with probability  $1 - \epsilon$  then there are decoders  $\Delta(\omega)$ such that given any  $x \in \ell_2^N$  we have with probability  $1 - 2\epsilon$ 

 $\|x - \Delta(\omega)\Phi(\omega)(x)\|_{\ell_2^N} \le C_0 \sigma_k(x)_{\ell_2^N}$ 

TAMU 2008 – p 12/2

- Instance-optimality in probability
- Theorem holds for Gaussian, Bernouli and other random families of matrices with  $\epsilon = e^{-cn}$
- Range of k is  $k \le c_0 n / \log(N/n)$

• If  $\{\Phi(\omega)\}$  satisfies RIP of order 3k and boundedness each with probability  $1 - \epsilon$  then there are decoders  $\Delta(\omega)$ such that given any  $x \in \ell_2^N$  we have with probability  $1 - 2\epsilon$ 

 $\|x - \Delta(\omega)\Phi(\omega)(x)\|_{\ell_2^N} \le C_0 \sigma_k(x)_{\ell_2^N}$ 

TAMU 2008 – p 12/2

- Instance-optimality in probability
- Theorem holds for Gaussian, Bernouli and other random families of matrices with  $\epsilon = e^{-cn}$
- Range of k is  $k \le c_0 n / \log(N/n)$
- Decoder is impractical

• If  $\{\Phi(\omega)\}$  satisfies RIP of order 3k and boundedness each with probability  $1 - \epsilon$  then there are decoders  $\Delta(\omega)$ such that given any  $x \in \ell_2^N$  we have with probability  $1 - 2\epsilon$ 

 $\|x - \Delta(\omega)\Phi(\omega)(x)\|_{\ell_2^N} \le C_0 \sigma_k(x)_{\ell_2^N}$ 

TAMU 2008 – p. 12/2<sup>,</sup>

- Instance-optimality in probability
- Theorem holds for Gaussian, Bernouli and other random families of matrices with  $\epsilon = e^{-cn}$
- Range of k is  $k \le c_0 n / \log(N/n)$
- Decoder is impractical
- Notice probability is on the draw of  $\Phi$  not on x

By far the most intriguing part of Compressed Sensing is the decoding - Hence this Workshop

- By far the most intriguing part of Compressed Sensing is the decoding - Hence this Workshop
- There are continuing debates as to which decoding is numerically fastest

- By far the most intriguing part of Compressed Sensing is the decoding - Hence this Workshop
- There are continuing debates as to which decoding is numerically fastest
- Some common decoders



- By far the most intriguing part of Compressed Sensing is the decoding - Hence this Workshop
- There are continuing debates as to which decoding is numerically fastest
- Some common decoders
  - $\ell_1$  minimization: Long history most everyone in this workshop has contributed to its development



- By far the most intriguing part of Compressed Sensing is the decoding - Hence this Workshop
- There are continuing debates as to which decoding is numerically fastest
- Some common decoders
  - $\ell_1$  minimization: Long history most everyone in this workshop has contributed to its development
  - Greedy algorithms find support of a good approximation vector and then decode using l<sub>2</sub> minimization (Gilbert-Tropp; Needell-Vershynin(ROMP), Donoho (STOMP)), Needell - Tropp (CoSAMP)

- By far the most intriguing part of Compressed Sensing is the decoding - Hence this Workshop
- There are continuing debates as to which decoding is numerically fastest
- Some common decoders
  - $\ell_1$  minimization: Long history most everyone in this workshop has contributed to its development
  - Greedy algorithms find support of a good approximation vector and then decode using l<sub>2</sub> minimization (Gilbert-Tropp; Needell-Vershynin(ROMP), Donoho (STOMP)), Needell - Tropp (CoSAMP)
  - Iterative Reweighted Least Squares (Osborne, Daubechies-DeVore-Fornasier-Gunturk)

TAMU 2008 – p. 13/2<sup>•</sup>

Range of Instance Optimality: When combined with encoder does it give full range of instance optimality?

- Range of Instance Optimality: When combined with encoder does it give full range of instance optimality?
- Number of computations to decode?



- Range of Instance Optimality: When combined with encoder does it give full range of instance optimality?
- Number of computations to decode?
- Robustness to noise?



- Range of Instance Optimality: When combined with encoder does it give full range of instance optimality?
- Number of computations to decode?
- Robustness to noise?
- Theorems versus numerical examples



- Range of Instance Optimality: When combined with encoder does it give full range of instance optimality?
- Number of computations to decode?
- Robustness to noise?
- Theorems versus numerical examples
- Instance Optimality in Probability

- Range of Instance Optimality: When combined with encoder does it give full range of instance optimality?
- Number of computations to decode?
- Robustness to noise?
- Theorems versus numerical examples
- Instance Optimality in Probability
- Given that we cant construct best encoding matrices it seems that the best results would correspond to random draws of matrices

- Range of Instance Optimality: When combined with encoder does it give full range of instance optimality?
- Number of computations to decode?
- Robustness to noise?
- Theorems versus numerical examples
- Instance Optimality in Probability
- Given that we cant construct best encoding matrices it seems that the best results would correspond to random draws of matrices
- We shall concentrate the rest of the talk on which decoders give instance optimality in  $\ell_2$  with high probability for a large range of k

Gilbert and Tropp proved that for Bernouli random matrices, OMP captures a k sparse vector with high probability for the large range of k

TAMU 2008 – p. 15/21

- Gilbert and Tropp proved that for Bernouli random matrices, OMP captures a k sparse vector with high probability for the large range of k
- Are there practical decoders that give instance optimality in  $\ell_2$ ?



- Gilbert and Tropp proved that for Bernouli random matrices, OMP captures a k sparse vector with high probability for the large range of k
- Are there practical decoders that give instance optimality in  $\ell_2$ ?
- Wojtaszczek has shown that for Gaussian matrices  $\ell_1$  minimization does the job for the large range of k



- Gilbert and Tropp proved that for Bernouli random matrices, OMP captures a k sparse vector with high probability for the large range of k
- Are there practical decoders that give instance optimality in  $\ell_2$ ?
- Wojtaszczek has shown that for Gaussian matrices  $\ell_1$  minimization does the job for the large range of k
- The proof is nontrivial and rests on a geometric property of Gaussian matrices: with high probability on the draw  $\Phi(\omega)$  the image of the unit  $\ell_1$  ball under such matrices will with high probability contain an  $\ell_2$  ball of radius  $\frac{c_0\sqrt{L}}{\sqrt{L}} = L := n/\log(N/n)$

radius 
$$\frac{c_0 \sqrt{L}}{\sqrt{n}}$$
,  $L := n/\log(N/n)$ 

- Gilbert and Tropp proved that for Bernouli random matrices, OMP captures a k sparse vector with high probability for the large range of k
- Are there practical decoders that give instance optimality in  $\ell_2$ ?
- Wojtaszczek has shown that for Gaussian matrices  $\ell_1$  minimization does the job for the large range of k
- The proof is nontrivial and rests on a geometric property of Gaussian matrices: with high probability on the draw  $\Phi(\omega)$  the image of the unit  $\ell_1$  ball under such matrices will with high probability contain an  $\ell_2$  ball of radius  $\frac{c_0\sqrt{L}}{\sqrt{n}}$ ,  $L := n/\log(N/n)$
- This Geometric Property does not hold for more general random families, e.g. Bernoulli.
  TAMU 2008 p. 15/2

### **Extension to more general families**

DeVore-Petrova-Wojtaszczek



### **Extension to more general families**

- DeVore-Petrova-Wojtaszczek
- We introduce new geometry to handle general random families (including Bernoulli)



### **Extension to more general families**

- DeVore-Petrova-Wojtaszczek
- We introduce new geometry to handle general random families (including Bernoulli)
- $||y||_J := \max(\{||y||_{\ell_{\infty}(\mathbb{R}^n)}, \sqrt{L}||y||_{\ell_2(\mathbb{R}^n)}\}, \quad L := \log(N/n).$

- DeVore-Petrova-Wojtaszczek
- We introduce new geometry to handle general random families (including Bernoulli)
- $\|y\|_J := \max(\left\{ \|y\|_{\ell_{\infty}(\mathbb{R}^n)}, \sqrt{L}\|y\|_{\ell_2(\mathbb{R}^n)} \right\}, \quad L := \log(N/n).$
- The unit ball  $U_J$  under this norm consists of all  $y \in \mathbb{R}^n$ such that  $\|y\|_{\ell_{\infty}} \leq 1/\sqrt{n}$  and  $\|y\|_{\ell_2} \leq \sqrt{L}/\sqrt{n}$

- DeVore-Petrova-Wojtaszczek
- We introduce new geometry to handle general random families (including Bernoulli)
- $\|y\|_J := \max(\left\{ \|y\|_{\ell_{\infty}(\mathbb{R}^n)}, \sqrt{L}\|y\|_{\ell_2(\mathbb{R}^n)} \right\}, \quad L := \log(N/n).$
- The unit ball  $U_J$  under this norm consists of all  $y \in \mathbb{R}^n$ such that  $\|y\|_{\ell_{\infty}} \leq 1/\sqrt{n}$  and  $\|y\|_{\ell_2} \leq \sqrt{L}/\sqrt{n}$

TAMU 2008 – p. 16/21

 $U_J$  is a trimmed  $\ell_2$  ball

- DeVore-Petrova-Wojtaszczek
- We introduce new geometry to handle general random families (including Bernoulli)
- $||y||_J := \max(\{||y||_{\ell_{\infty}(\mathbb{R}^n)}, \sqrt{L}||y||_{\ell_2(\mathbb{R}^n)}\}, \quad L := \log(N/n).$
- The unit ball  $U_J$  under this norm consists of all  $y \in \mathbb{R}^n$ such that  $\|y\|_{\ell_{\infty}} \leq 1/\sqrt{n}$  and  $\|y\|_{\ell_2} \leq \sqrt{L}/\sqrt{n}$
- $U_J$  is a trimmed  $\ell_2$  ball
- With high probability on the draw  $\Phi(\omega)$  we have  $\Phi(U(\ell_1^N)) \supset c_0 U_J$

- DeVore-Petrova-Wojtaszczek
- We introduce new geometry to handle general random families (including Bernoulli)
- $||y||_J := \max(\{||y||_{\ell_{\infty}(\mathbb{R}^n)}, \sqrt{L}||y||_{\ell_2(\mathbb{R}^n)}\}, \quad L := \log(N/n).$
- The unit ball  $U_J$  under this norm consists of all  $y \in \mathbb{R}^n$ such that  $\|y\|_{\ell_{\infty}} \leq 1/\sqrt{n}$  and  $\|y\|_{\ell_2} \leq \sqrt{L}/\sqrt{n}$
- $U_J$  is a trimmed  $\ell_2$  ball
- With high probability on the draw  $\Phi(\omega)$  we have  $\Phi(U(\ell_1^N)) \supset c_0 U_J$
- Using this result we can prove that encoding with Bernouli (and more general random families) decoding with  $\ell_1$  minimization is instance optimal for the large \_\_\_\_\_\_ range of k \_\_\_\_\_\_\_\_

• Let  $\Phi(\omega)$  be a random family of  $n \times N$  compressed sensing matrices

- Let  $\Phi(\omega)$  be a random family of  $n \times N$  compressed sensing matrices
- We suppose n = am with a and m integers



- Let  $\Phi(\omega)$  be a random family of  $n \times N$  compressed sensing matrices
- We suppose n = am with a and m integers
- Let  $\Phi_j := [\phi_1^j, \dots, \phi_N^1]$ ,  $j = 1, \dots, a$ , be obtained from the rows  $i = jm + 1, \dots, j(m + 1)$  of  $\Phi$

- Let  $\Phi(\omega)$  be a random family of  $n \times N$  compressed sensing matrices
- We suppose n = am with a and m integers
- Let  $\Phi_j := [\phi_1^j, \dots, \phi_N^1]$ ,  $j = 1, \dots, a$ , be obtained from the rows  $i = jm + 1, \dots, j(m + 1)$  of  $\Phi$
- We shall give an iterative thresholding algorithm for decoding

- Let  $\Phi(\omega)$  be a random family of  $n \times N$  compressed sensing matrices
- We suppose n = am with a and m integers
- Let  $\Phi_j := [\phi_1^j, \dots, \phi_N^1]$ ,  $j = 1, \dots, a$ , be obtained from the rows  $i = jm + 1, \dots, j(m + 1)$  of  $\Phi$
- We shall give an iterative thresholding algorithm for decoding
- At each step it will choose a set  $\overline{\Lambda}_j$  of coordinates (of x) where inner products with residuals are large

- Let  $\Phi(\omega)$  be a random family of  $n \times N$  compressed sensing matrices
- We suppose n = am with a and m integers
- Let  $\Phi_j := [\phi_1^j, \dots, \phi_N^1]$ ,  $j = 1, \dots, a$ , be obtained from the rows  $i = jm + 1, \dots, j(m + 1)$  of  $\Phi$
- We shall give an iterative thresholding algorithm for decoding
- At each step it will choose a set  $\overline{\Lambda}_j$  of coordinates (of x) where inner products with residuals are large
- These coordinates are added to all previously selected coordinates to give the new set  $\Lambda_j := \overline{\Lambda}_1 \cup \cdots \cup \overline{\Lambda}_j$  which is our current set of coordinates where we think x is big

- Let  $\Phi(\omega)$  be a random family of  $n \times N$  compressed sensing matrices
- We suppose n = am with a and m integers
- Let  $\Phi_j := [\phi_1^j, \dots, \phi_N^1]$ ,  $j = 1, \dots, a$ , be obtained from the rows  $i = jm + 1, \dots, j(m + 1)$  of  $\Phi$
- We shall give an iterative thresholding algorithm for decoding
- At each step it will choose a set  $\overline{\Lambda}_j$  of coordinates (of x) where inner products with residuals are large
- These coordinates are added to all previously selected coordinates to give the new set  $\Lambda_j := \overline{\Lambda}_1 \cup \cdots \cup \overline{\Lambda}_j$  which is our current set of coordinates where we think x is big
- There is a threshold parameter  $\delta > 0$  which can be chosen for example as 1/8 TAMU 2008 – p. 17/2

• Fix a value of k



- Fix a value of k
- Initial Step: Let  $r^1 := y^1 = \Phi_1(x)$  be the initial residual

TAMU 2008 – p. 18/2<sup>,</sup>

- Fix a value of k
- Initial Step: Let  $r^1 := y^1 = \Phi_1(x)$  be the initial residual
- $\bar{\Lambda}_1 := \Lambda_1$  is defined as the set of all coordinates  $\nu$  where  $|\langle y^1, \phi_{\nu}^1 \rangle| > \delta k^{-1/2} ||y^1||_{\ell_2}$

- Fix a value of k
- Initial Step: Let  $r^1 := y^1 = \Phi_1(x)$  be the initial residual
- $\bar{\Lambda}_1 := \Lambda_1$  is defined as the set of all coordinates  $\nu$  where  $|\langle y^1, \phi_{\nu}^1 \rangle| > \delta k^{-1/2} ||y^1||_{\ell_2}$
- Define  $x^1$  as the least squares minimizer

$$x^{1} := \underset{\sup(z) \subset \Lambda_{1}}{\operatorname{Argmin}} \|y^{1} - \Phi^{1}z\|_{\ell^{2}}$$

- Fix a value of k
- Initial Step: Let  $r^1 := y^1 = \Phi_1(x)$  be the initial residual
- $\bar{\Lambda}_1 := \Lambda_1$  is defined as the set of all coordinates  $\nu$  where  $|\langle y^1, \phi_{\nu}^1 \rangle| > \delta k^{-1/2} ||y^1||_{\ell_2}$

TAMU 2008 – p. 18/2′

Define  $x^1$  as the least squares minimizer

$$x^{1} := \underset{\sup(z) \subset \Lambda_{1}}{\operatorname{Argmin}} \|y^{1} - \Phi^{1}z\|_{\ell^{2}}$$

• New residual  $r^1 := \Phi_2(x - x^1) = y^2 - \Phi_2(x^1)$ 

- Fix a value of k
- Initial Step: Let  $r^1 := y^1 = \Phi_1(x)$  be the initial residual
- $\bar{\Lambda}_1 := \Lambda_1$  is defined as the set of all coordinates  $\nu$  where  $|\langle y^1, \phi_{\nu}^1 \rangle| > \delta k^{-1/2} ||y^1||_{\ell_2}$
- Define  $x^1$  as the least squares minimizer

$$x^{1} := \underset{\sup(z) \subset \Lambda_{1}}{\operatorname{Argmin}} \|y^{1} - \Phi^{1}z\|_{\ell^{2}}$$

- New residual  $r^1 := \Phi_2(x x^1) = y^2 \Phi_2(x^1)$
- General Step: Repeat the above with  $r^1$  replaced by  $r^j$  and  $\Phi_1$  replaced by  $\Phi_j$

TAMU 2008 – p. 18/2′

• We do not want sets with cardinality larger than 2k

- We do not want sets with cardinality larger than 2k
- If the size of  $\Lambda_j := \overline{\Lambda}_1 \cup \ldots \cup \overline{\Lambda}_j$  exceeds 2k trim back to 2k coordinates by removing elements from  $\overline{\Lambda}_j$

TAMU 2008 – p. 19/21

- We do not want sets with cardinality larger than 2k
- If the size of  $\Lambda_j := \overline{\Lambda}_1 \cup \ldots \cup \overline{\Lambda}_j$  exceeds 2k trim back to 2k coordinates by removing elements from  $\overline{\Lambda}_j$

TAMU 2008 – p. 19/21

• This gives the output set  $\Lambda$ 

- We do not want sets with cardinality larger than 2k
- If the size of  $\Lambda_j := \overline{\Lambda}_1 \cup \ldots \cup \overline{\Lambda}_j$  exceeds 2k trim back to 2k coordinates by removing elements from  $\overline{\Lambda}_j$
- This gives the output set  $\Lambda$
- $\bar{x}$  is the least squares solution for this  $\Lambda$  using  $y^j$  and  $\Phi_j$

- We do not want sets with cardinality larger than 2k
- If the size of  $\Lambda_j := \overline{\Lambda}_1 \cup \ldots \cup \overline{\Lambda}_j$  exceeds 2k trim back to 2k coordinates by removing elements from  $\overline{\Lambda}_j$
- This gives the output set  $\Lambda$
- $\bar{x}$  is the least squares solution for this  $\Lambda$  using  $y^j$  and  $\Phi_j$
- otherwise stop at step j = a and output  $\overline{x} := x^a$

The next slide will give an instance optimality for this algorithm



- The next slide will give an instance optimality for this algorithm
- Here are the assumptions on the random family  $\Phi(\omega)$ ,  $\omega \in \Omega$  for this theorem to hold

TAMU 2008 – p. 20/2<sup>,</sup>

- The next slide will give an instance optimality for this algorithm
- Here are the assumptions on the random family  $\Phi(\omega)$ ,  $\omega \in \Omega$  for this theorem to hold

TAMU 2008 – p. 20/2<sup>,</sup>

• The entries of  $\Phi$  are independent

- The next slide will give an instance optimality for this algorithm
- Here are the assumptions on the random family  $\Phi(\omega)$ ,  $\omega \in \Omega$  for this theorem to hold
  - The entries of  $\Phi$  are independent
  - Given  $\delta > 0$  and  $x \in \mathbb{R}^N$  with probability  $\geq 1 C_1 e^{-c_1 n \delta^2}$  we have

 $|\|\Phi(x)\|_{\ell_2} - \|x\|_{\ell_2}| \le \delta \|x\|_{\ell_2}$ 



- The next slide will give an instance optimality for this algorithm
- Here are the assumptions on the random family  $\Phi(\omega)$ ,  $\omega \in \Omega$  for this theorem to hold
  - The entries of  $\Phi$  are independent
  - Given  $\delta > 0$  and  $x \in \mathbb{R}^N$  with probability  $\geq 1 C_1 e^{-c_1 n \delta^2}$  we have

 $|||\Phi(x)||_{\ell_2} - ||x||_{\ell_2}| \le \delta ||x||_{\ell_2}$ 

• Given  $\delta > 0$ ,  $l \in \{1, ..., N\}$  and  $z \in \mathbb{R}^n$  with probability  $\geq 1 - C_1 e^{-c_1 n \delta^2}$  we have

 $|\langle z, \phi_l \rangle| \le \delta \|z\|_{\ell_2}$ 

- The next slide will give an instance optimality for this algorithm
- Here are the assumptions on the random family  $\Phi(\omega)$ ,  $\omega \in \Omega$  for this theorem to hold
  - The entries of  $\Phi$  are independent
  - Given  $\delta > 0$  and  $x \in \mathbb{R}^N$  with probability  $\geq 1 C_1 e^{-c_1 n \delta^2}$  we have

 $|||\Phi(x)||_{\ell_2} - ||x||_{\ell_2}| \le \delta ||x||_{\ell_2}$ 

• Given  $\delta > 0$ ,  $l \in \{1, ..., N\}$  and  $z \in \mathbb{R}^n$  with probability  $\geq 1 - C_1 e^{-c_1 n \delta^2}$  we have

 $|\langle z, \phi_l \rangle| \le \delta \|z\|_{\ell_2}$ 

TAMU 2008 – p. 20/2<sup>,</sup>

These properties hold for Gaussian and Bernouli

## **Thresholding Theorem**

THEOREM: Given any  $0 < \delta \leq 1/8\sqrt{3}$  and given any r, s > 0. The thresholding decoder applied with this choice of  $\delta$  where the random matrices are of size  $n \times N$  with n = am and  $a := \lceil r \log N \rceil$  gives the following. For any  $x \in \mathbb{R}^N$ , with probability  $\geq 1 - n^{-s}$  the decoded vector  $\overline{x}$  of the above greedy decoder satisfies

 $||x - \bar{x}||_{\ell_2} \le n^{-r} ||x|| + 198\sigma_k(x),$ 

TAMU 2008 – p. 21/2′

for any  $k \leq c(\delta, s)n/(\log N)^2$ .

• Loss of a logarithm in the range of k

## **Thresholding Theorem**

THEOREM: Given any  $0 < \delta \leq 1/8\sqrt{3}$  and given any r, s > 0. The thresholding decoder applied with this choice of  $\delta$  where the random matrices are of size  $n \times N$  with n = am and  $a := \lceil r \log N \rceil$  gives the following. For any  $x \in \mathbb{R}^N$ , with probability  $\geq 1 - n^{-s}$  the decoded vector  $\overline{x}$  of the above greedy decoder satisfies

 $||x - \bar{x}||_{\ell_2} \le n^{-r} ||x|| + 198\sigma_k(x),$ 

TAMU 2008 – p. 21/2′

for any  $k \leq c(\delta, s)n/(\log N)^2$ .

- Loss of a logarithm in the range of k
- Do not quite have instance optimality

## **Thresholding Theorem**

THEOREM: Given any  $0 < \delta \leq 1/8\sqrt{3}$  and given any r, s > 0. The thresholding decoder applied with this choice of  $\delta$  where the random matrices are of size  $n \times N$  with n = am and  $a := \lceil r \log N \rceil$  gives the following. For any  $x \in \mathbb{R}^N$ , with probability  $\geq 1 - n^{-s}$  the decoded vector  $\overline{x}$  of the above greedy decoder satisfies

 $||x - \bar{x}||_{\ell_2} \le n^{-r} ||x|| + 198\sigma_k(x),$ 

TAMU 2008 — p. 21/2

for any  $k \leq c(\delta, s)n/(\log N)^2$ .

- Loss of a logarithm in the range of k
- Do not quite have instance optimality
- Fine for any numerical purpose