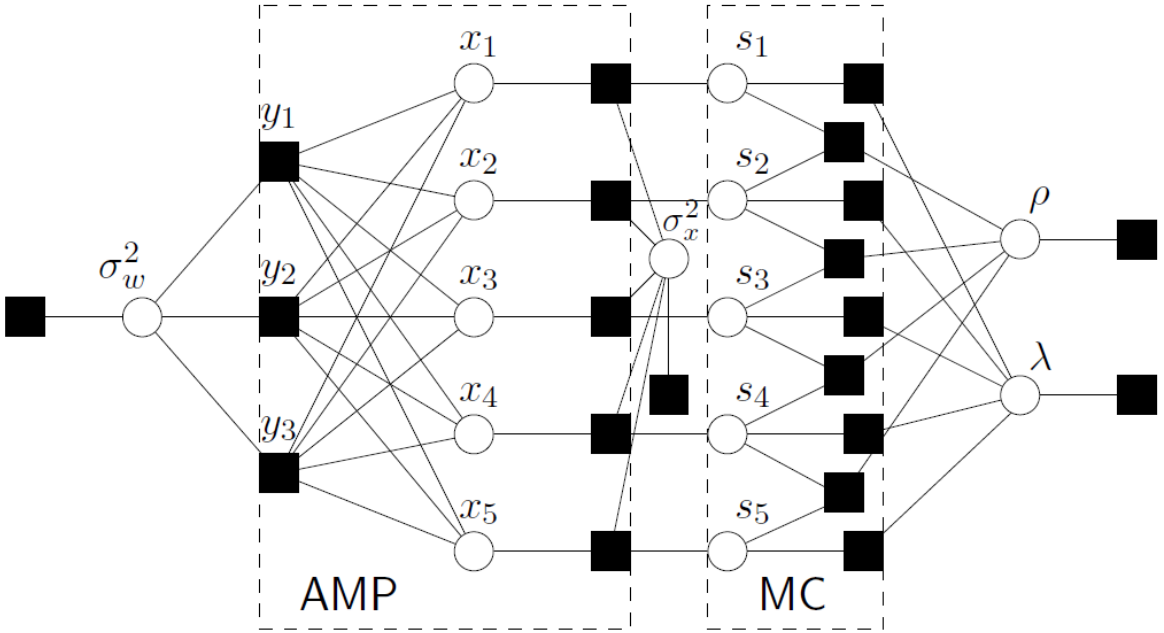
Linear Inverse Problems

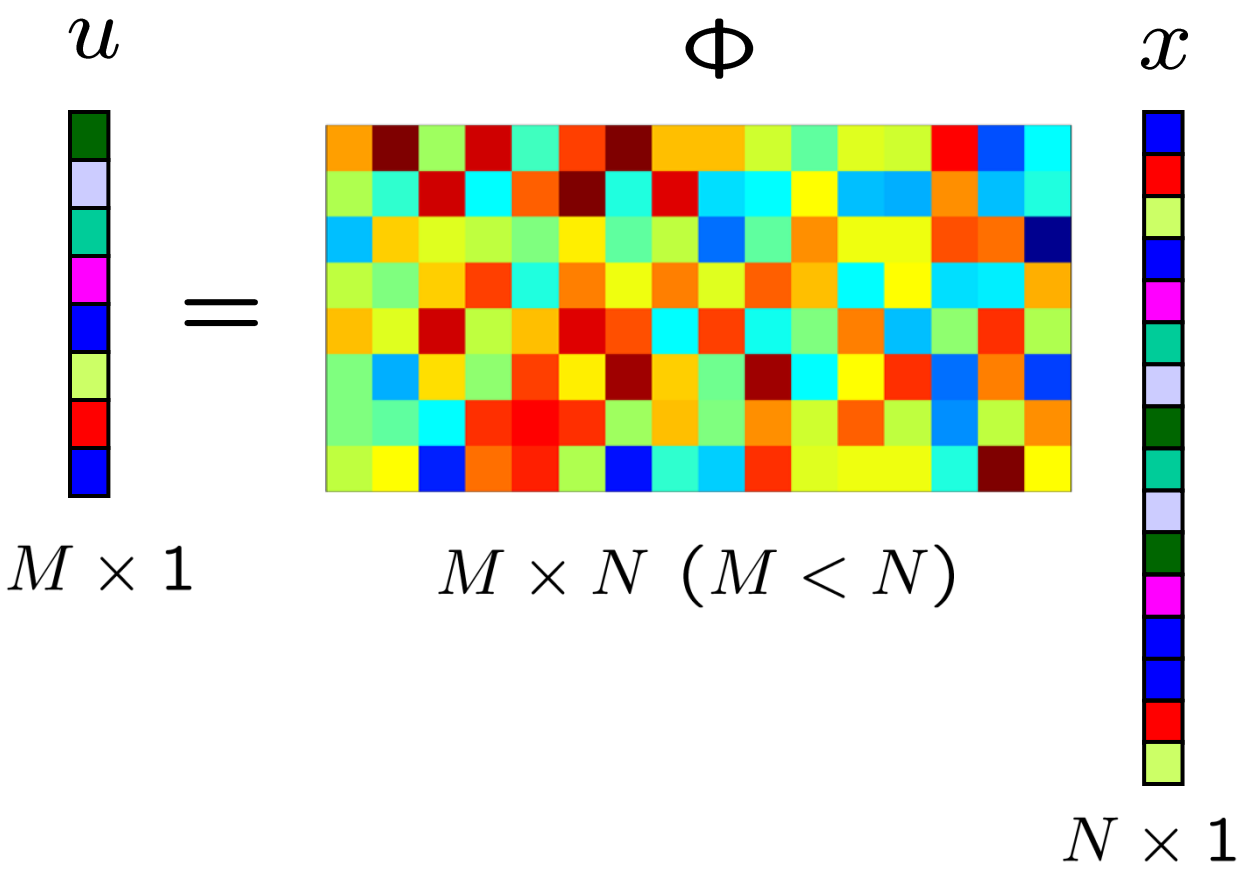
with a Compressive Sensing Flavor



Volkan Cevher
Matthias Seeger

Probabilistic Graphical Models—Fall 2011

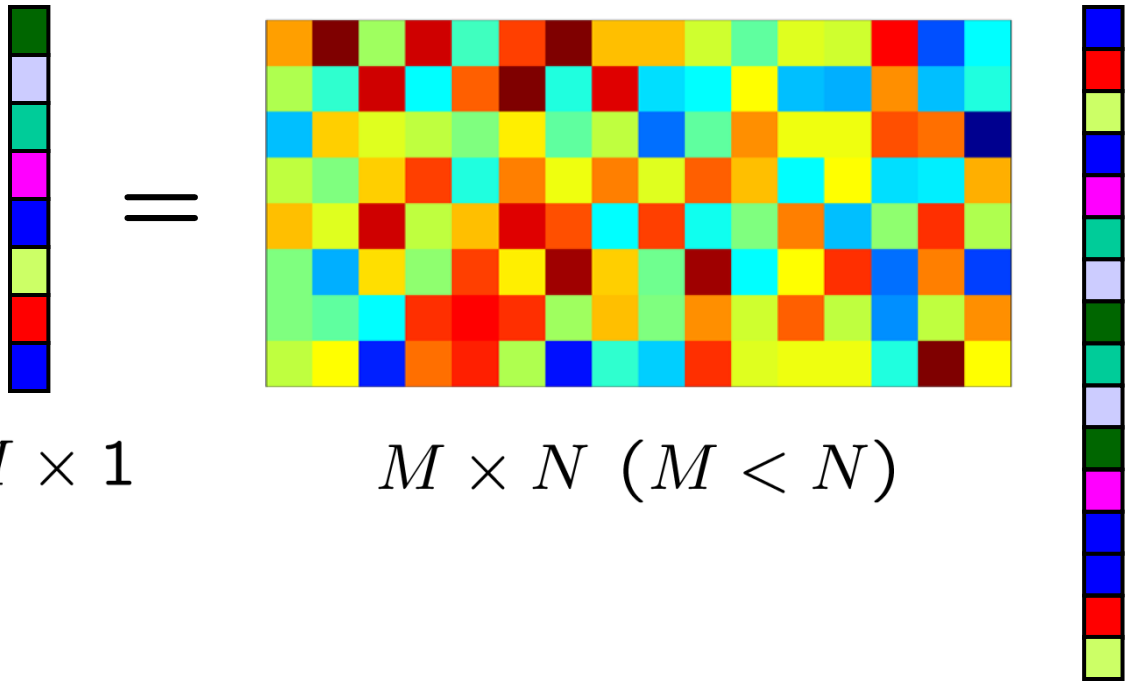
Linear Inverse Problems



compressive sensing	non-adaptive measurements
machine learning	dictionary of features
communications	MIMO user detection
theoretical computer science	sketching matrix / expander

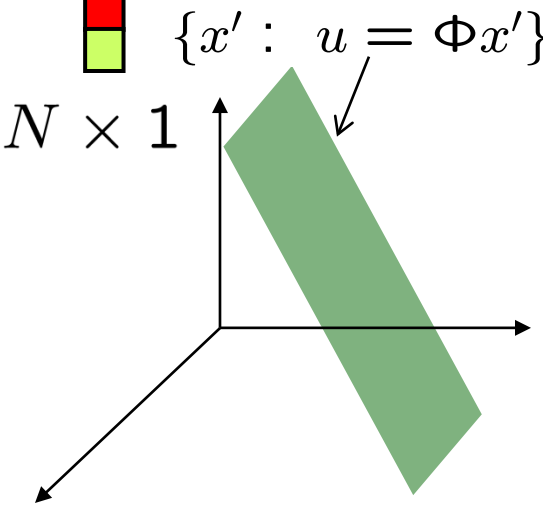
Linear Inverse Problems

u $\quad$ Φ $\quad$ x



$M \times 1$ $\quad$ $M \times N \ (M < N)$

$N \times 1$



- **Challenge:** Null space of Φ: $\mathcal{N}(\Phi)$

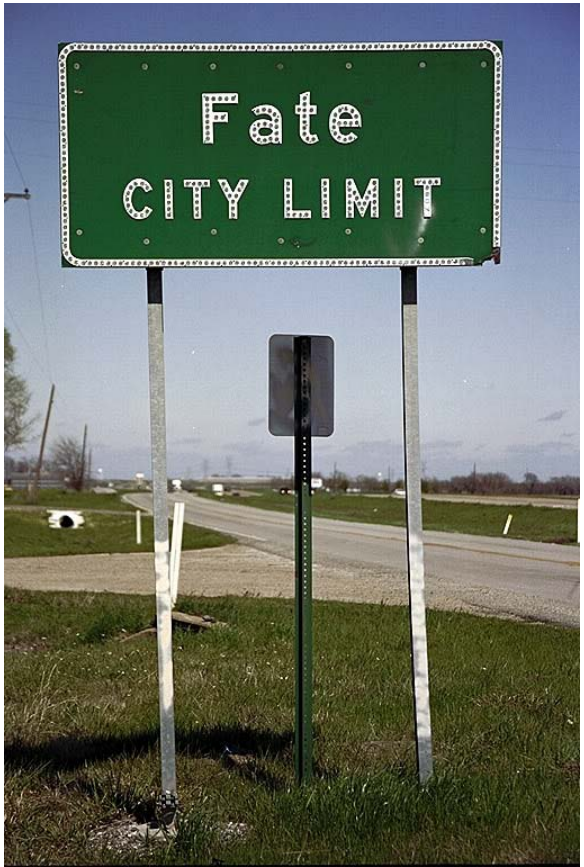
$$x' = x + v \Rightarrow u, \ \forall v \in \mathcal{N}(\Phi)$$

Approaches

	Deterministic	**Probabilistic**
Prior	 parsity	$f(x)$
Metric	ℓ_p-norm*	likelihood/ posterior

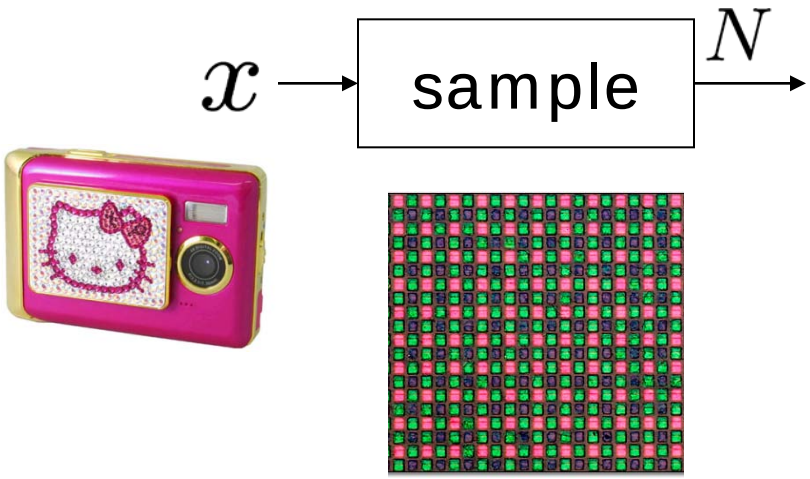
$$* : \|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$$

A Deterministic View
(Motivation)



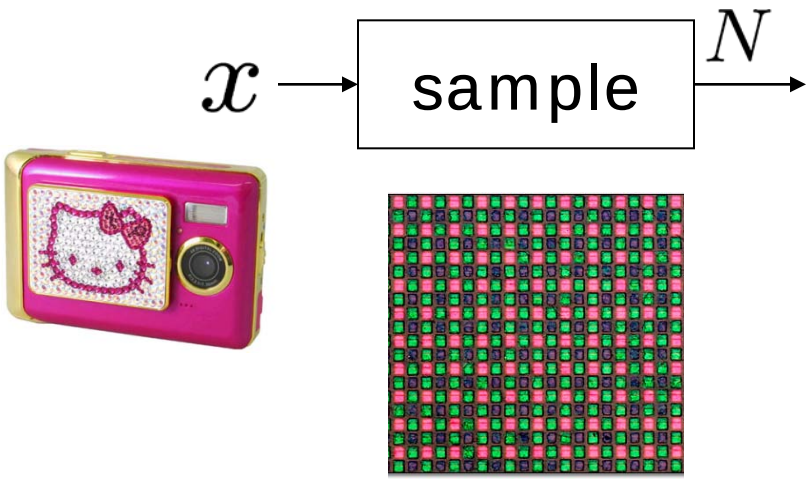
Sensing by *Sampling*

- Long-established paradigm for digital data acquisition
  - uniformly ***sample*** data at Nyquist rate (2x Fourier bandwidth)



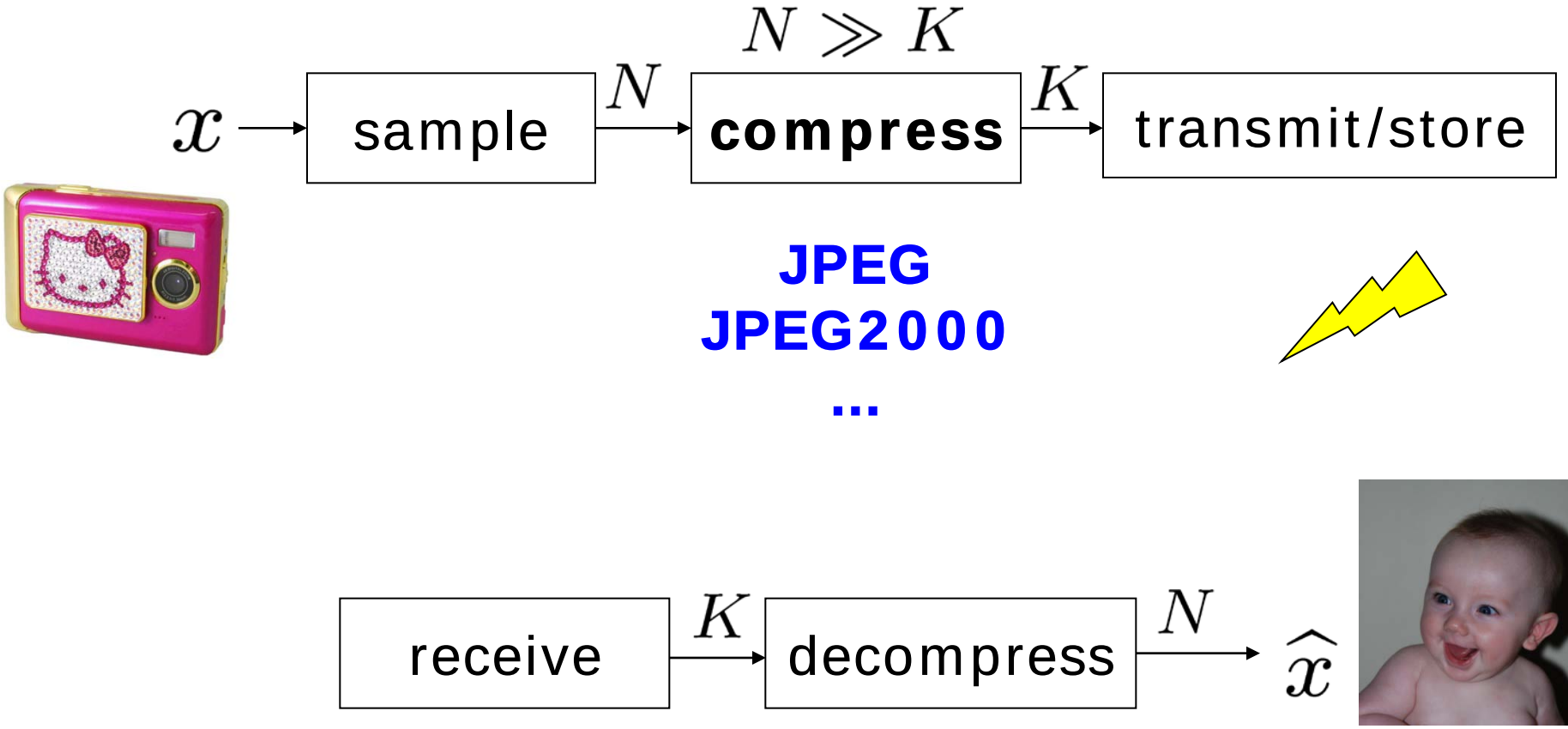
Sensing by *Sampling*

- Long-established paradigm for digital data acquisition
  - uniformly ***sample*** data at Nyquist rate (2x Fourier bandwidth)



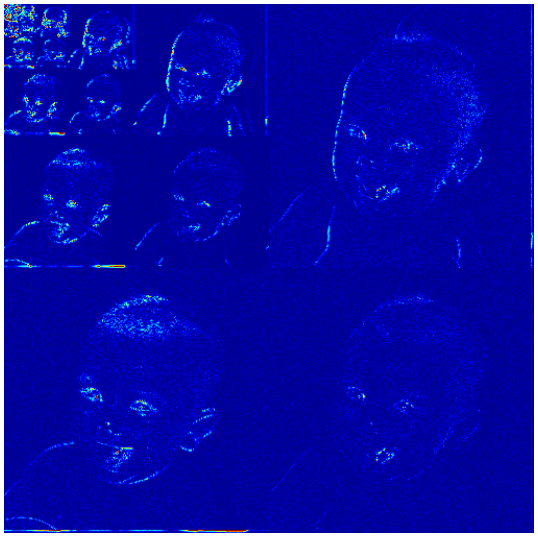
Sensing by *Sampling*

- Long-established paradigm for digital data acquisition
  - uniformly ***sample*** data at Nyquist rate (2x Fourier bandwidth)
  - ***compress*** data



Sparsity / Compressibility

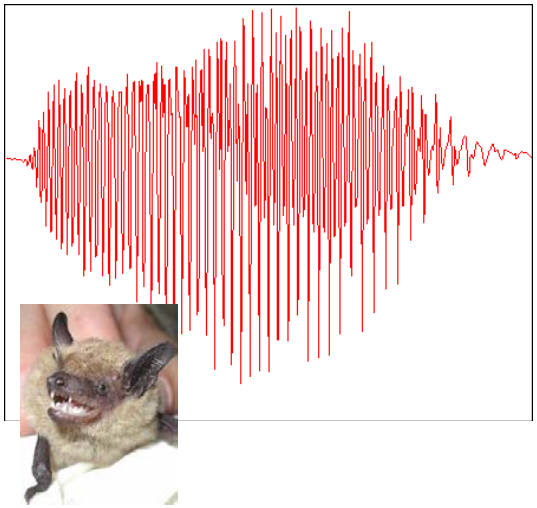
N
pixels

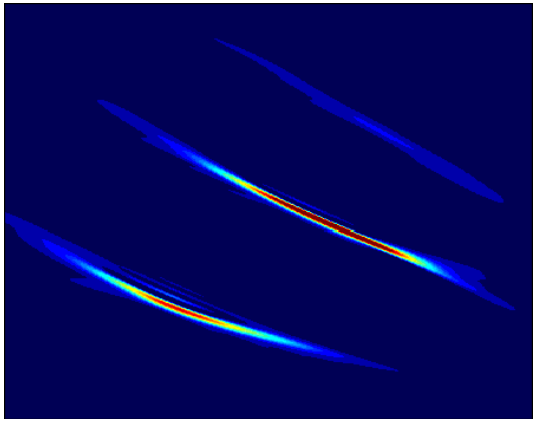


$K \ll N$
large
wavelet
coefficients

(blue = 0)

N
wideband
signal
samples

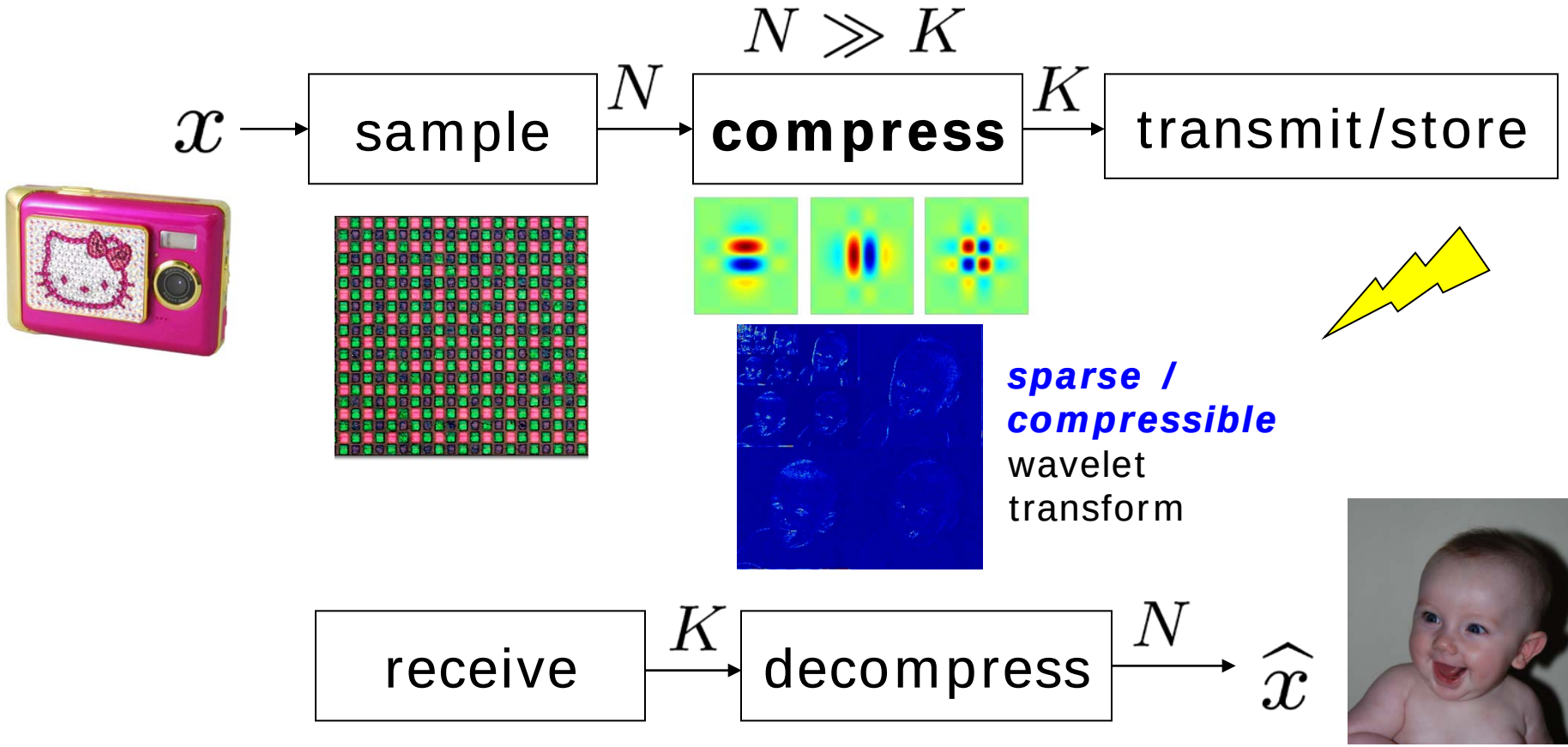




$K \ll N$
large
Gabor (TF)
coefficients

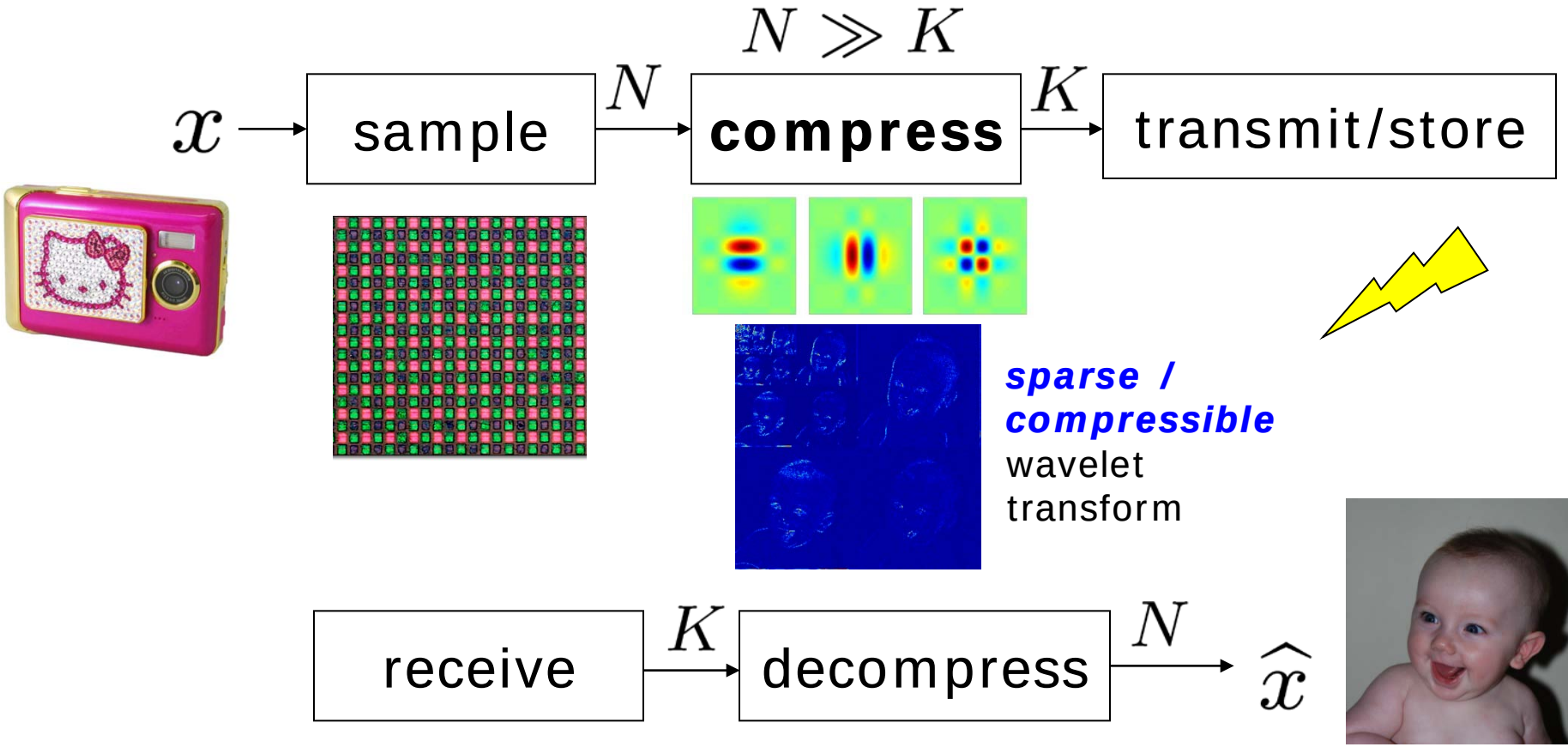
Sample / Compress

- Long-established paradigm for digital data acquisition
  - uniformly ***sample*** data at Nyquist rate
  - ***compress*** data

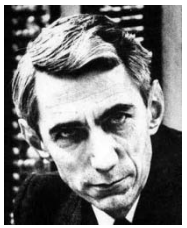


What's Wrong with this Picture?

- ***Why go to all the work to acquire*** *N* ***samples only to discard all but*** *K* ***pieces of data?***



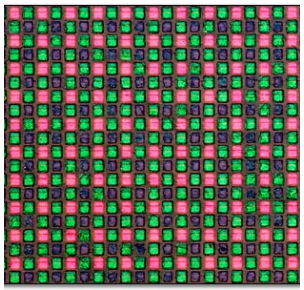
What's Wrong with this Picture?

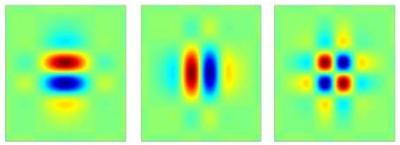


linear processing
linear signal model
(bandlimited subspace)

nonlinear processing
nonlinear signal model
(union of subspaces)

x → sample → N → **compress** → K → transmit/store





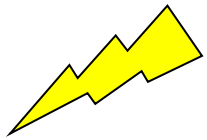
sparse / compressible
wavelet
transform

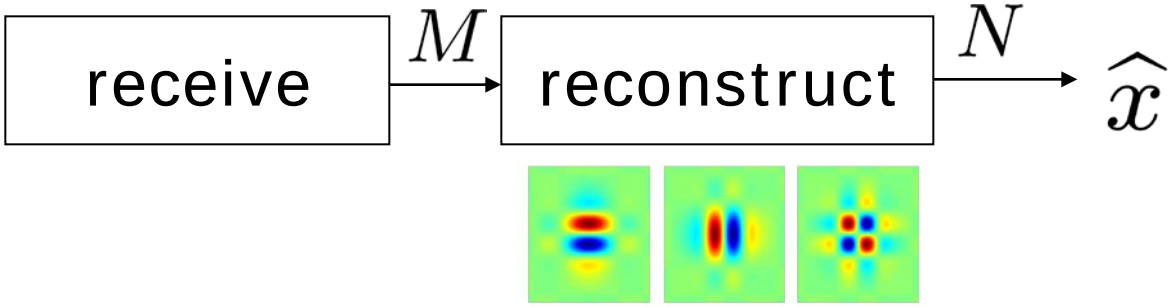
receive → K → decompress → N → $\hat{x}$

Compressive Sensing

- Directly acquire "***compressed***" data
- Replace samples by more general "measurements"

$$K \approx M \ll N$$



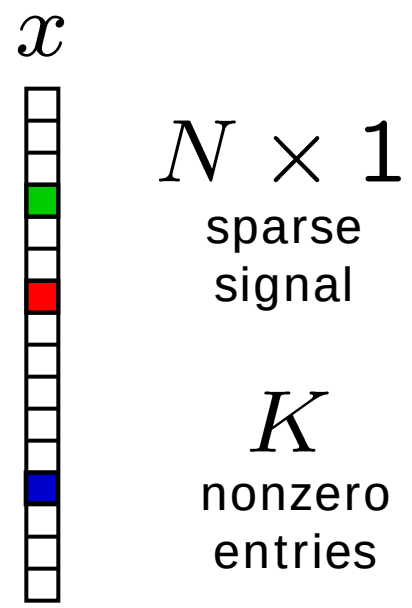


Linear Inverse Problems with CS Flavor

Theory I
Geometrical Perspective

Sampling

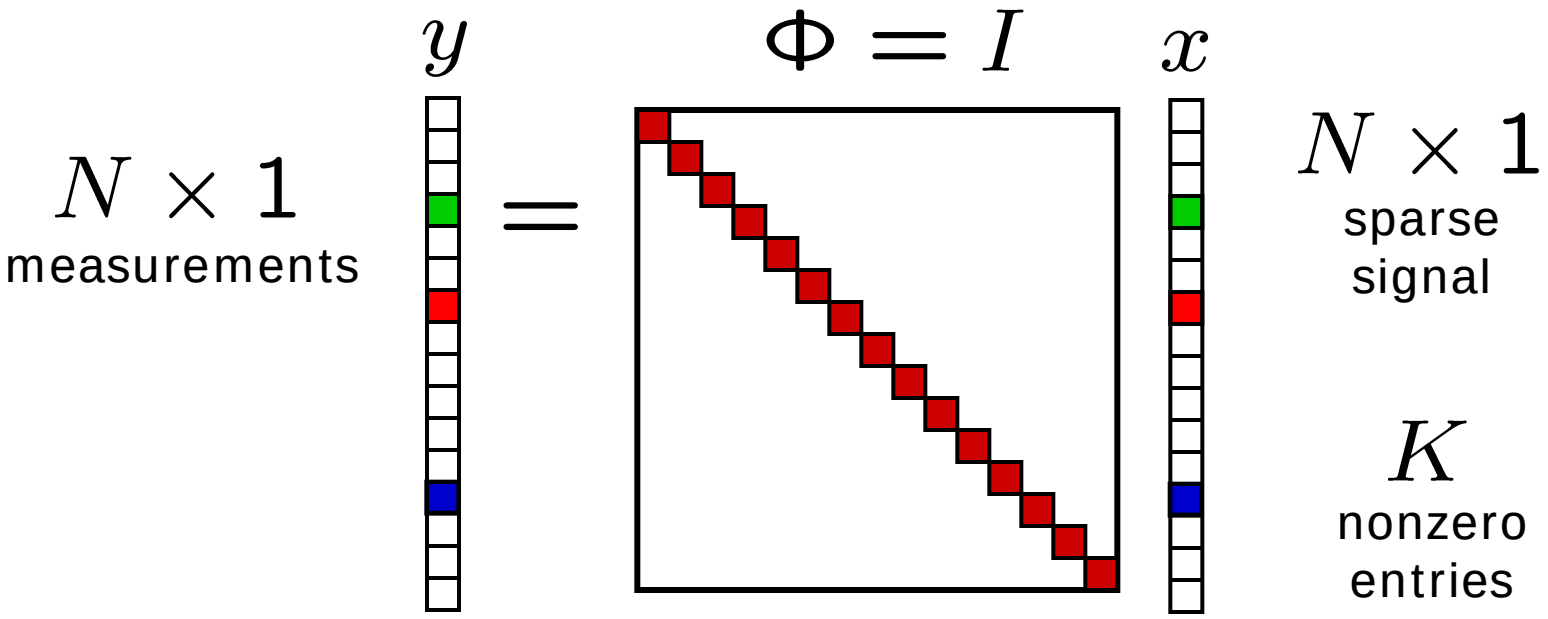
- Signal x is K-*sparse* in basis/dictionary Ψ
  - WLOG assume sparse in space domain $\Psi = I$



Sampling

- Signal x is K-*sparse* in basis/dictionary Ψ
  - WLOG assume sparse in space domain $\Psi = I$

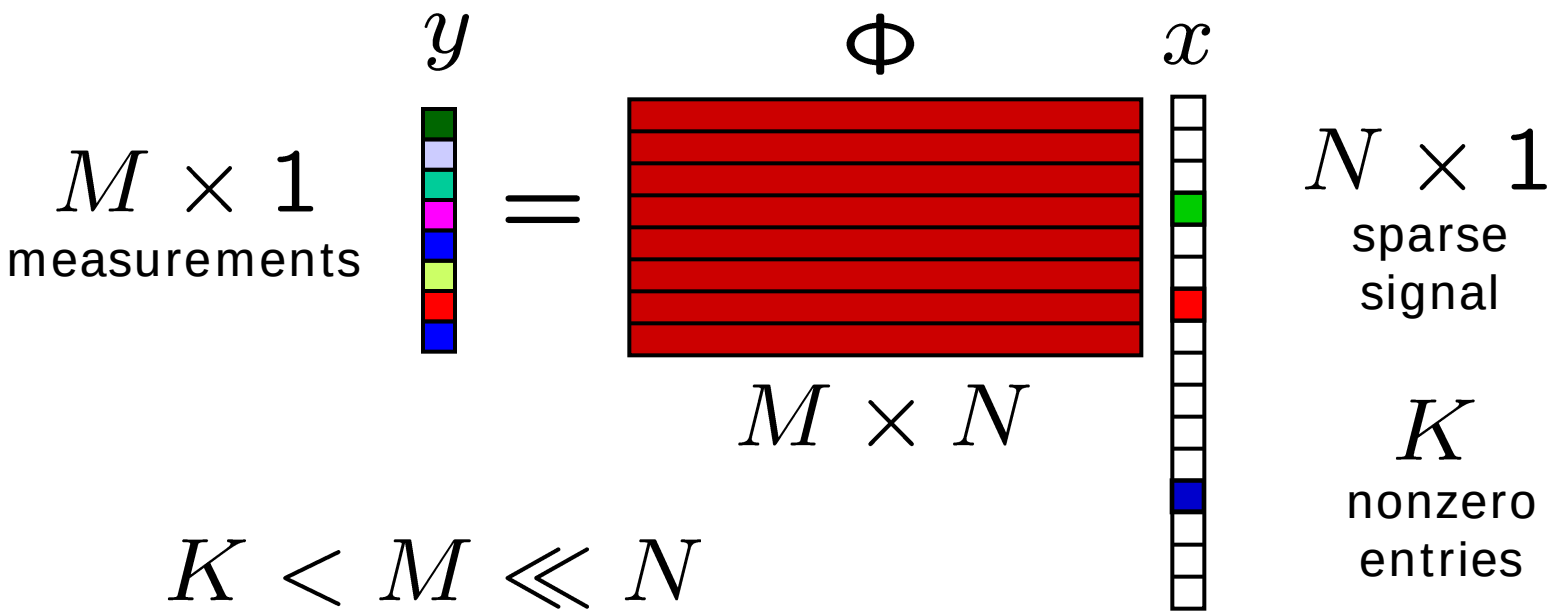
- **Samples**



Compressive Sampling

- When data is sparse/compressible, can directly acquire a ***condensed representation*** with no/little information loss through linear ***dimensionality reduction***

$$y = \Phi x$$

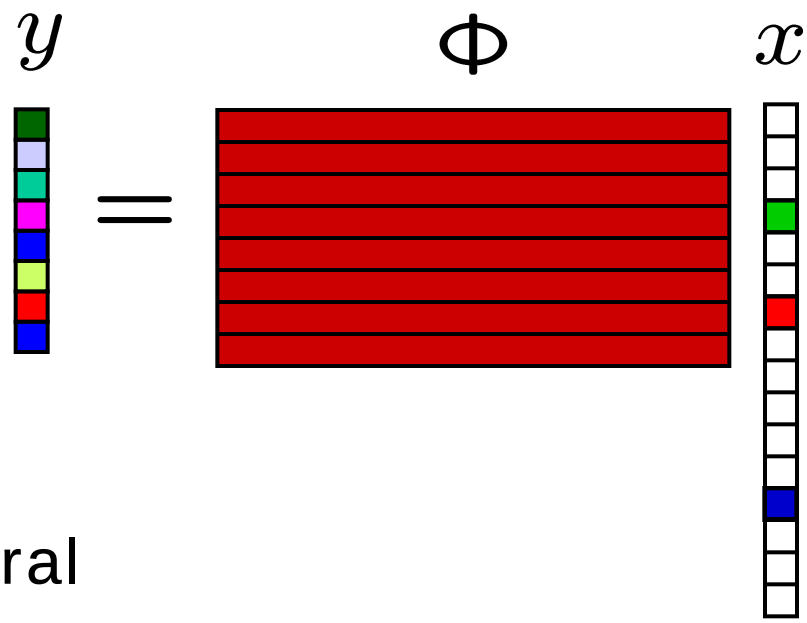


How Can It Work?

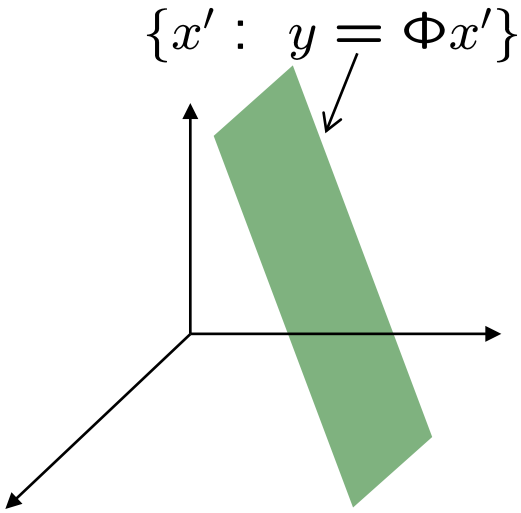
- Projection Φ **not full rank**...

  $M < N$

  ... and so **loses information** in general



- Ex: Infinitely many x's map to the same y

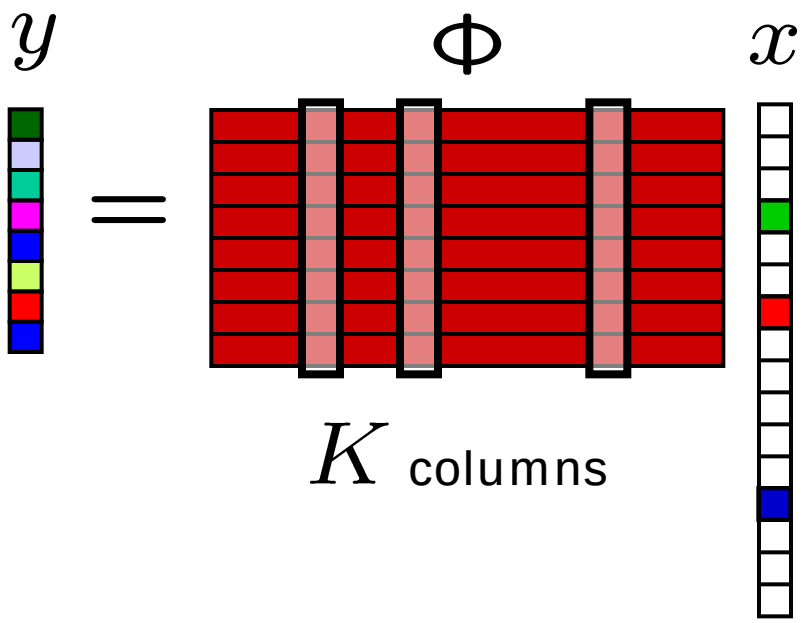


How Can It Work?

- Projection Φ not full rank...

  $M < N$

  ... and so loses information in general



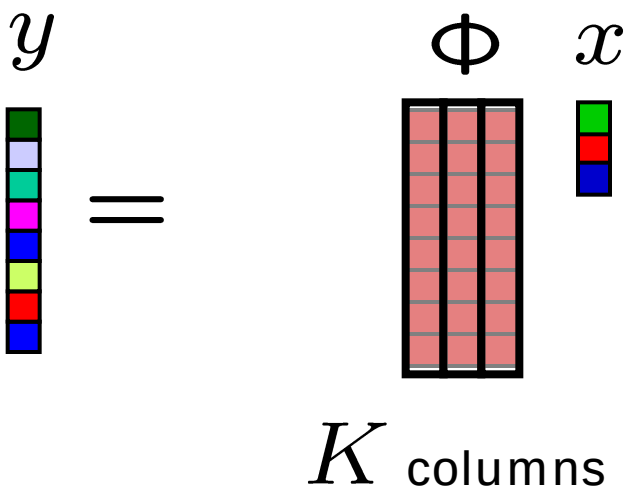
- But we are only interested in ***sparse*** vectors x

How Can It Work?

- Projection Φ not full rank...

  $M < N$

  ... and so loses information in general



- But we are only interested in ***sparse*** vectors

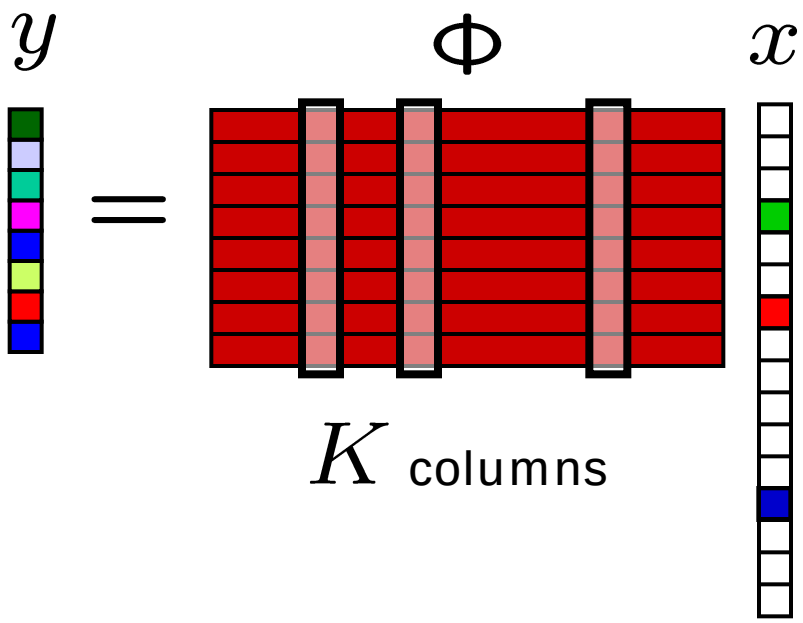
- Φ is effectively *M*x*K*

How Can It Work?

- Projection Φ not full rank...

  $M < N$

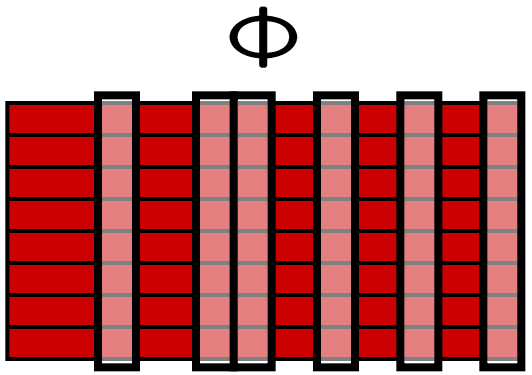
  ... and so loses information in general



- But we are only interested in ***sparse*** vectors

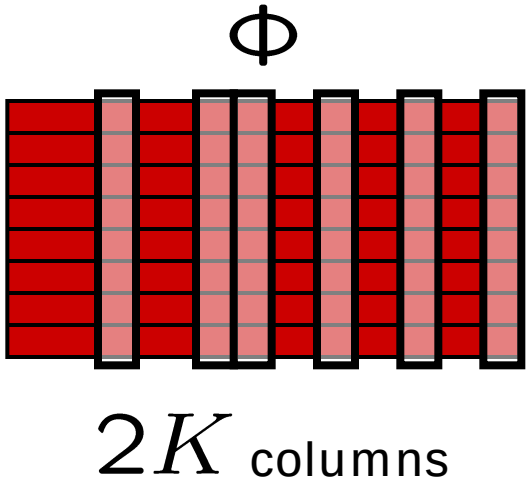
- **Design** Φ so that each of its *M*x*K* submatrices are full rank

How Can It Work?



- **Goal:** Design Φ so that its *M*x2*K* submatrices are full rank
  - difference $x_1 - x_2$ between two *K*-sparse vectors is 2*K* sparse in general
  - preserve information in *K*-sparse signals
  - **Restricted Isometry Property** (RIP) of order 2*K*

Unfortunately…

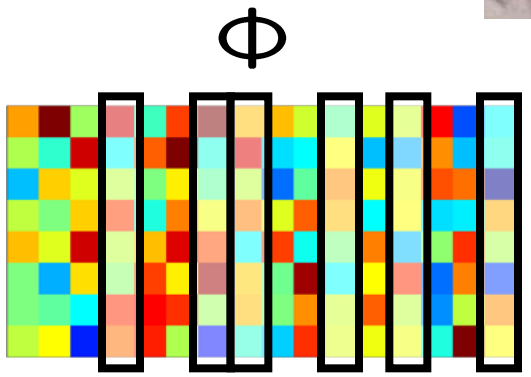


- **Goal:** Design Φ so that its *M*x2*K* submatrices are full rank (Restricted Isometry Property – RIP)

- Unfortunately, a combinatorial, **NP-complete design problem**

Insight from the 80's [Kashin, Gluskin]

- Draw Φ at **random**
  - iid Gaussian
  - iid Bernoulli ± 1
  - ...

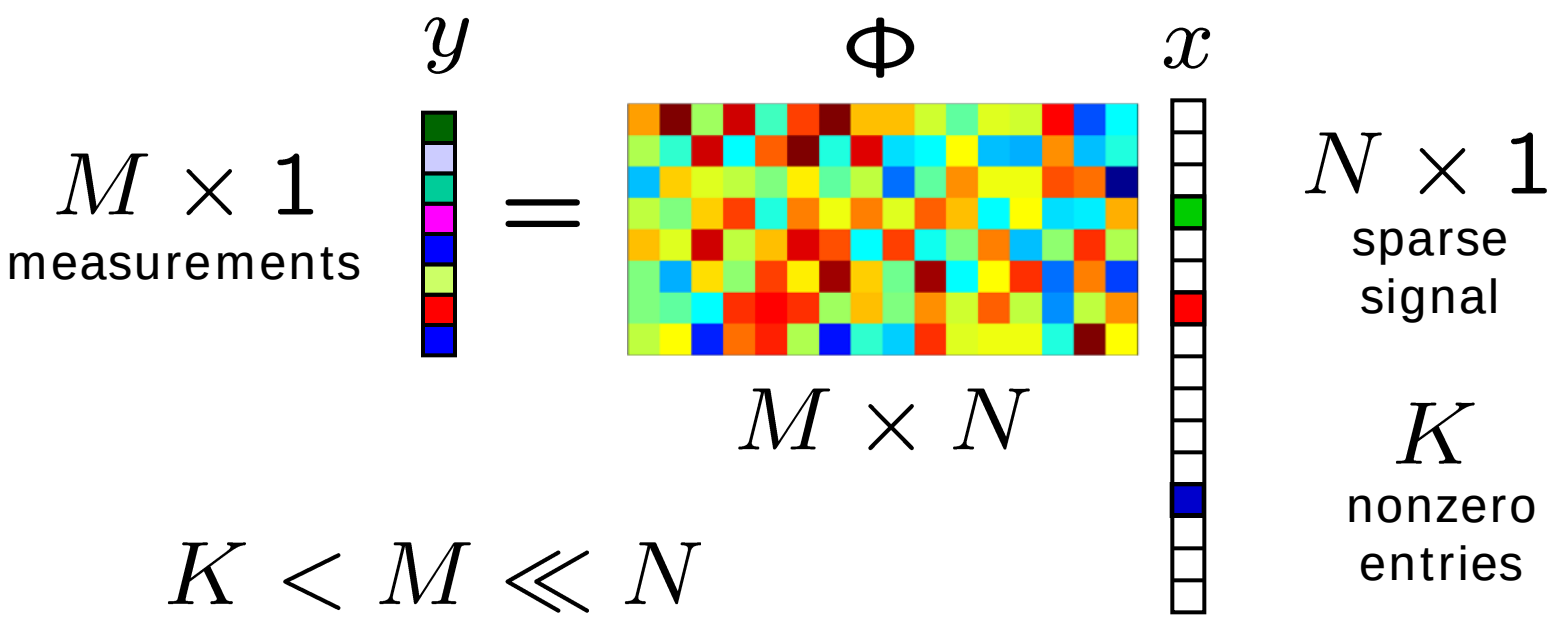


$2K$ columns

- Then Φ has the RIP with high probability as long as $M = O(K \log(N/K)) \ll N$
  - *M*x2*K* submatrices are full rank
  - stable embedding for sparse signals
  - extends to compressible signals

Compressive Data Acquisition

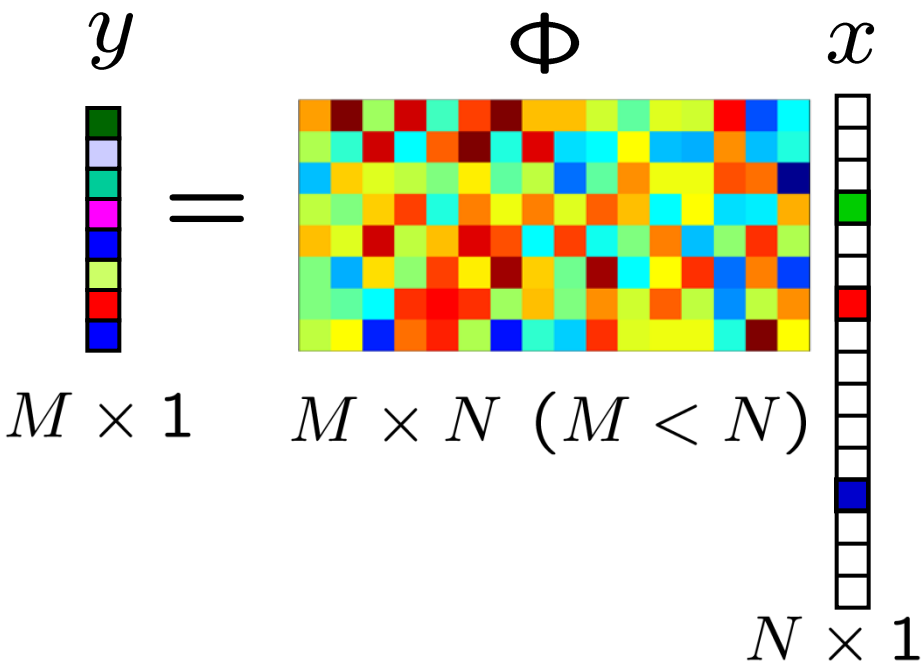
- Measurements y = **random linear combinations** of the entries of x
- WHP does not distort structure of sparse signals
  - no information loss



Compressive Sensing Recovery

1. Sparse / compressible x

   not sufficient alone



2. Projection Φ

   information preserving

   (restricted isometry property - RIP)

3. Decoding algorithms

   tractable

Compressive Sensing Recovery

- Recovery:
  (ill-posed inverse problem)

given $y = \Phi x$
find x (sparse)

- ℓ_2 **fast**

$$\widehat{x} = \arg\min_{y=\Phi x} \|x\|_2$$

$$\widehat{x} = (\Phi^T \Phi)^{-1} \Phi^T y$$

pseudoinverse

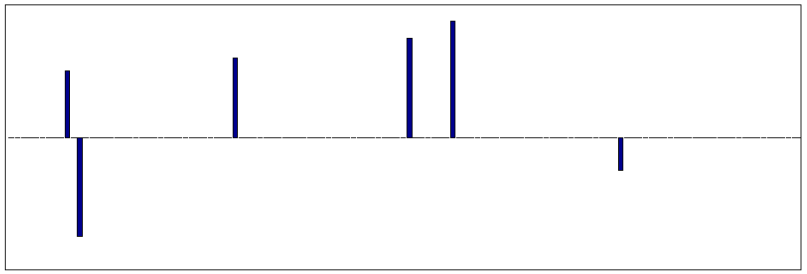
Compressive Sensing Recovery

- Recovery:
  (ill-posed inverse problem)

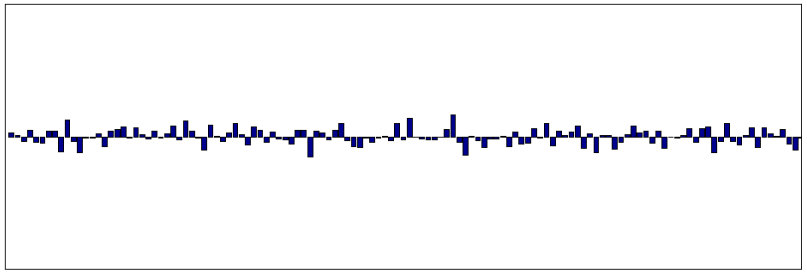
given $y = \Phi x$
find x (sparse)

- ℓ_2 **fast, wrong**

$$\hat{x} = \arg\min_{y=\Phi x} \|x\|_2$$



x



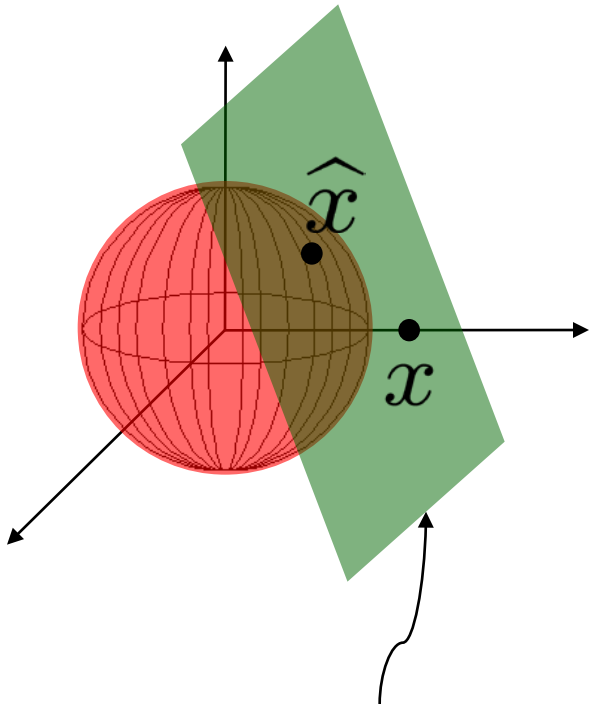
$$\hat{x} = (\Phi^T \Phi)^{-1} \Phi^T y$$

pseudoinverse

Why ℓ_2 Doesn't Work

for signals sparse in the
space / time domain

$$\widehat{x} = \arg\min_{y=\Phi x'} \|x'\|_2$$



least squares,
minimum ℓ_2 solution
is almost **never sparse**

null space of Φ
translated to x
(random angle)

Compressive Sensing Recovery

- Reconstruction/decoding: given $y = \Phi x$
  (ill-posed inverse problem) find x

- ℓ_2 fast, wrong $\quad \widehat{x} = \arg\min_{y=\Phi x} \|x\|_2$

- ℓ_0 $\quad \widehat{x} = \arg\min_{y=\Phi x} \|x\|_0$

number of nonzero entries

*"find **sparsest** x in translated nullspace"*

Compressive Sensing Recovery

- Reconstruction/decoding: given $y = \Phi x$
  (ill-posed inverse problem) find x

- ℓ_2 fast, wrong $\quad \hat{x} = \arg\min_{y=\Phi x} \|x\|_2$

- ℓ_0 **correct:** only $M = 2K$ measurements required to reconstruct K-sparse signal $\quad \hat{x} = \arg\min_{y=\Phi x} \|x\|_0$

number of nonzero entries

Compressive Sensing Recovery

- Reconstruction/decoding: given $y = \Phi x$
  (ill-posed inverse problem) find x

- ℓ_2 fast, wrong $\quad \hat{x} = \arg\min_{y=\Phi x} \|x\|_2$

- ℓ_0 **correct:**
  only $M=2K$ measurements required to reconstruct K-sparse signal

  slow: NP-hard algorithm

$$\hat{x} = \arg\min_{y=\Phi x} \|x\|_0$$

number of nonzero entries

Compressive Sensing Recovery

- Recovery: (ill-posed inverse problem)
  given $y = \Phi x$
  find x (sparse)

- ℓ_2 fast, wrong
  $\hat{x} = \arg\min_{y=\Phi x} \|x\|_2$

- ℓ_0 correct, slow
  $\hat{x} = \arg\min_{y=\Phi x} \|x\|_0$

- ℓ_1 **correct, efficient**
  mild oversampling
  [Candes, Romberg, Tao; Donoho]
  $\hat{x} = \arg\min_{y=\Phi x} \|x\|_1$
  linear program

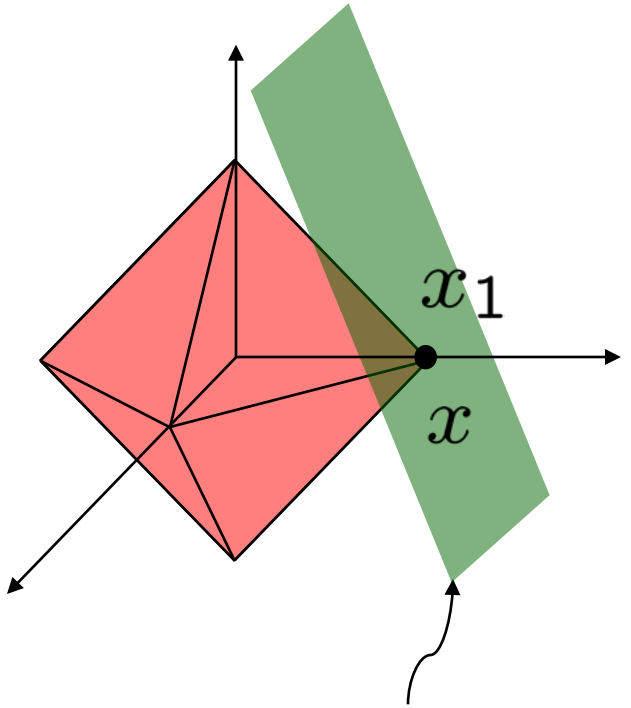
number of measurements required

$$M = O(K \log(N/K)) \ll N$$

Why ℓ_1 Works

for signals sparse in the **space / time domain**

$$\widehat{x} = \arg \min_{y=\Phi x'} \|x'\|_1$$



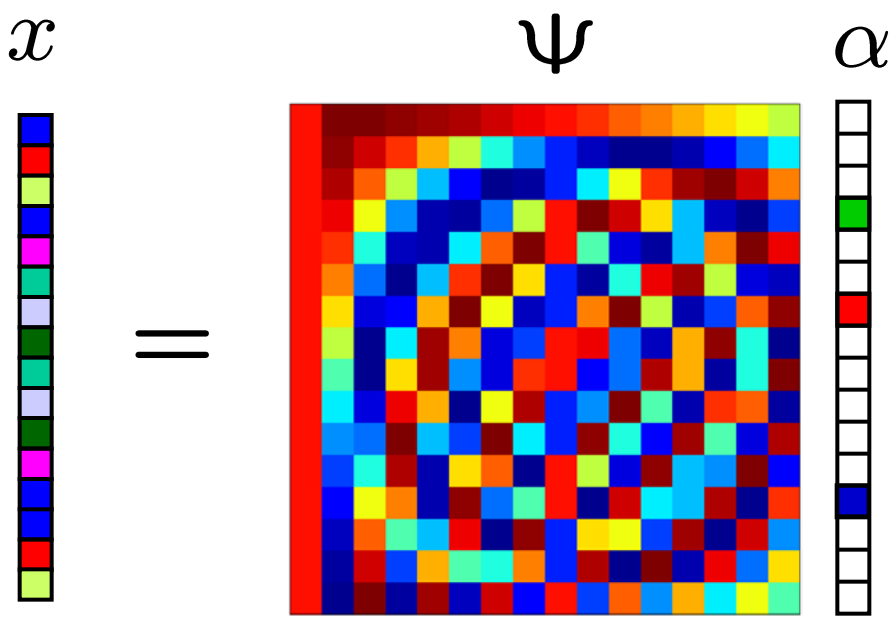
minimum ℓ_1 solution
= sparsest solution
(with high probability) if

$$M = O(K \log(N/K)) \ll N$$

Universality

- Random measurements can be used for signals sparse in *any* basis

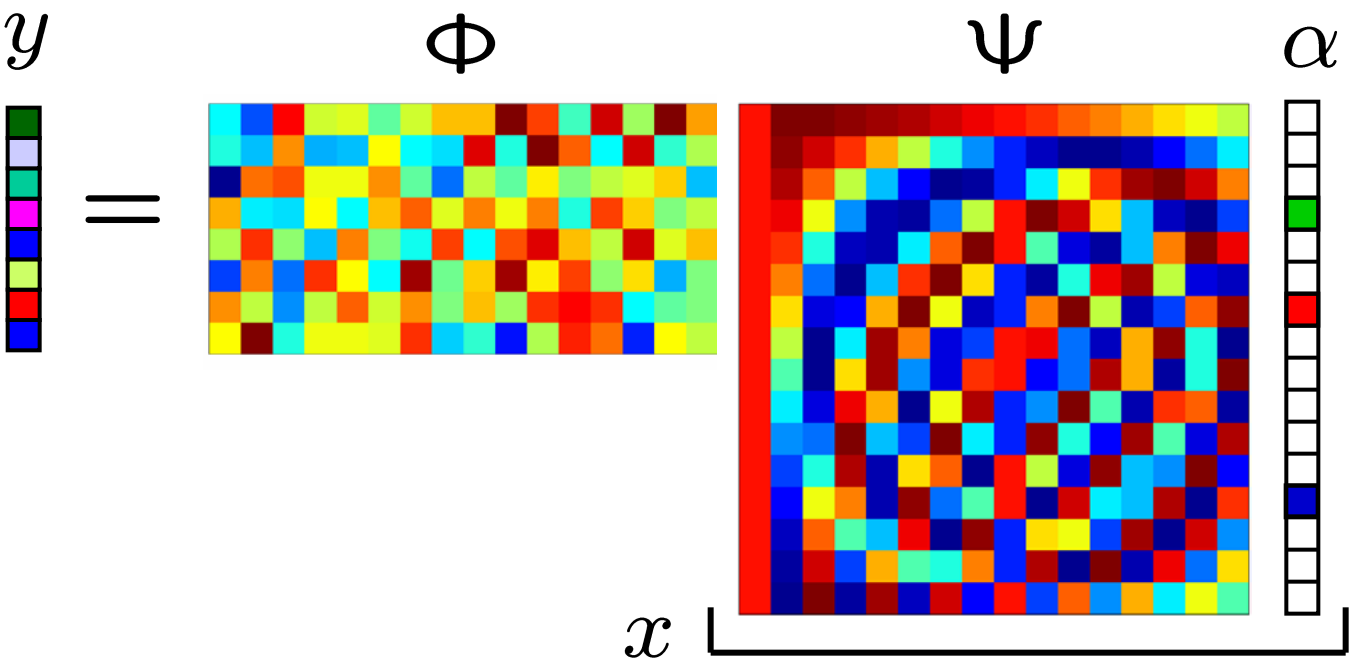
$$x = \Psi \alpha$$



Universality

- Random measurements can be used for signals sparse in *any* basis

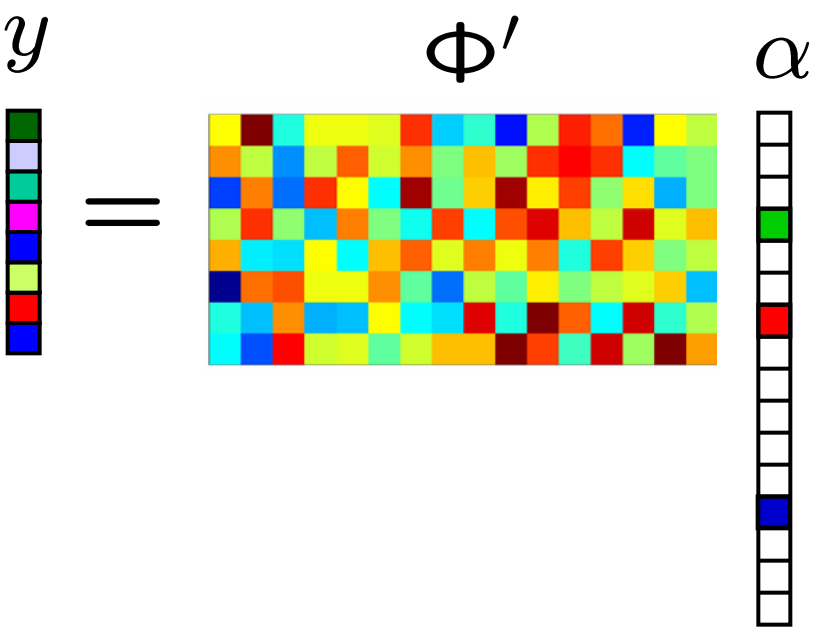
$$y = \Phi x = \Phi \Psi \alpha$$



Universality

- Random measurements can be used for signals sparse in *any* basis

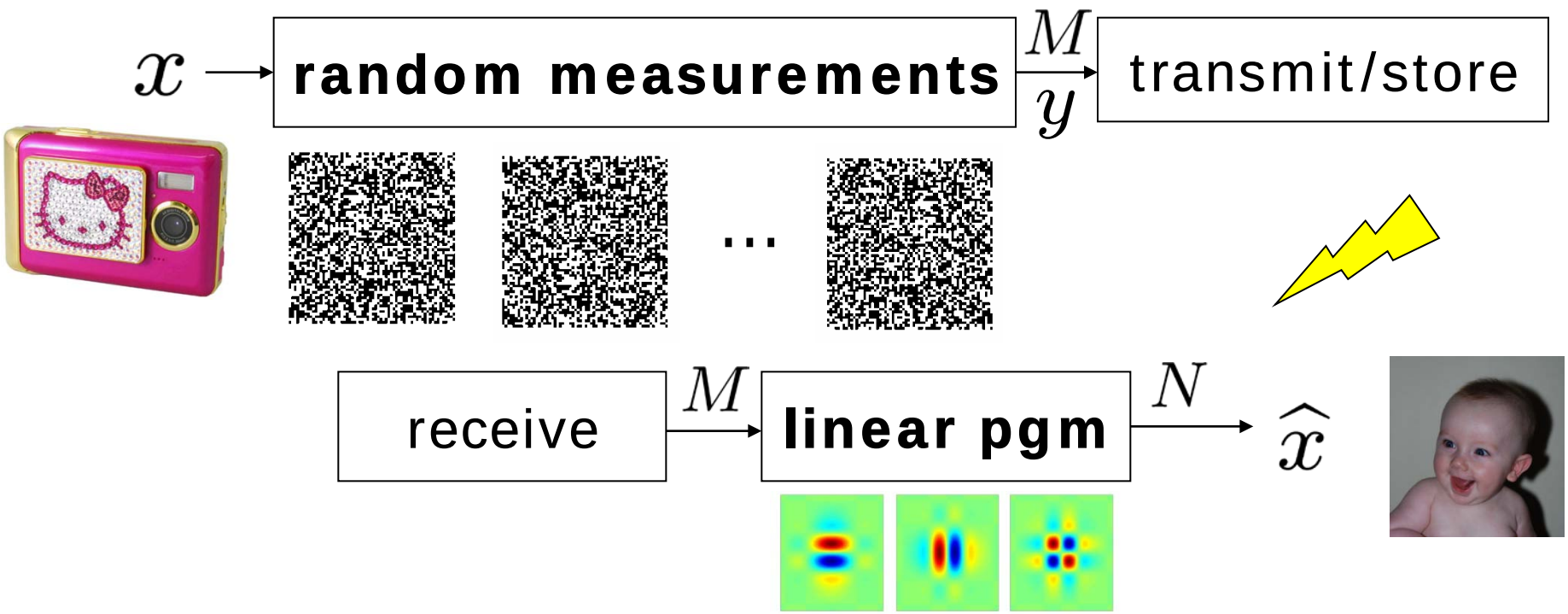
$$y = \Phi x = \Phi \Psi \alpha = \Phi' \alpha$$



Compressive Sensing

- Directly acquire "***compressed***" data
- Replace *N* samples by *M* random projections

$$M = O(K \log(N/K))$$

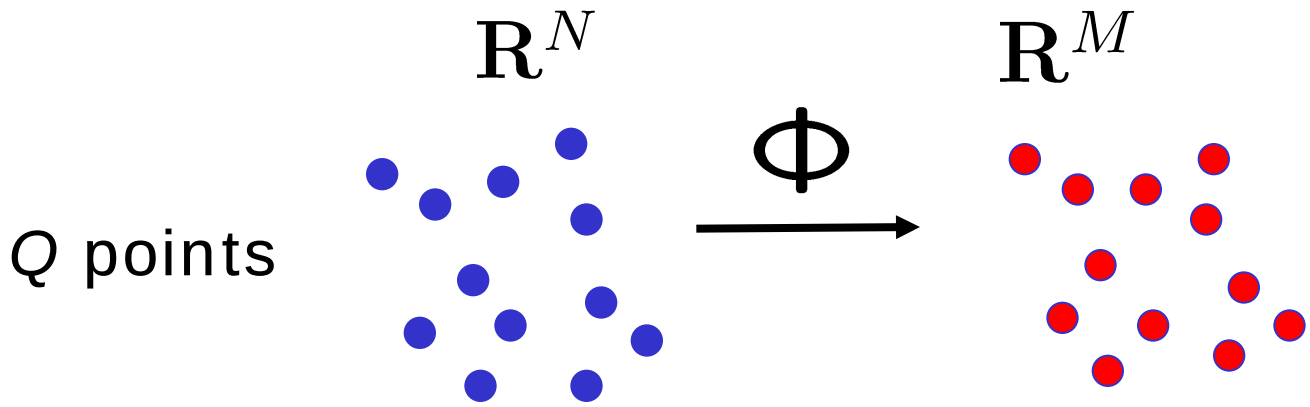


Linear Inverse Problems

Theory II
Stable Embedding

Johnson-Lindenstrauss Lemma

- JL Lemma: random projection stably embeds a cloud of Q points whp provided $M = O(\log Q)$



- Proved via concentration inequality (part of the homework)

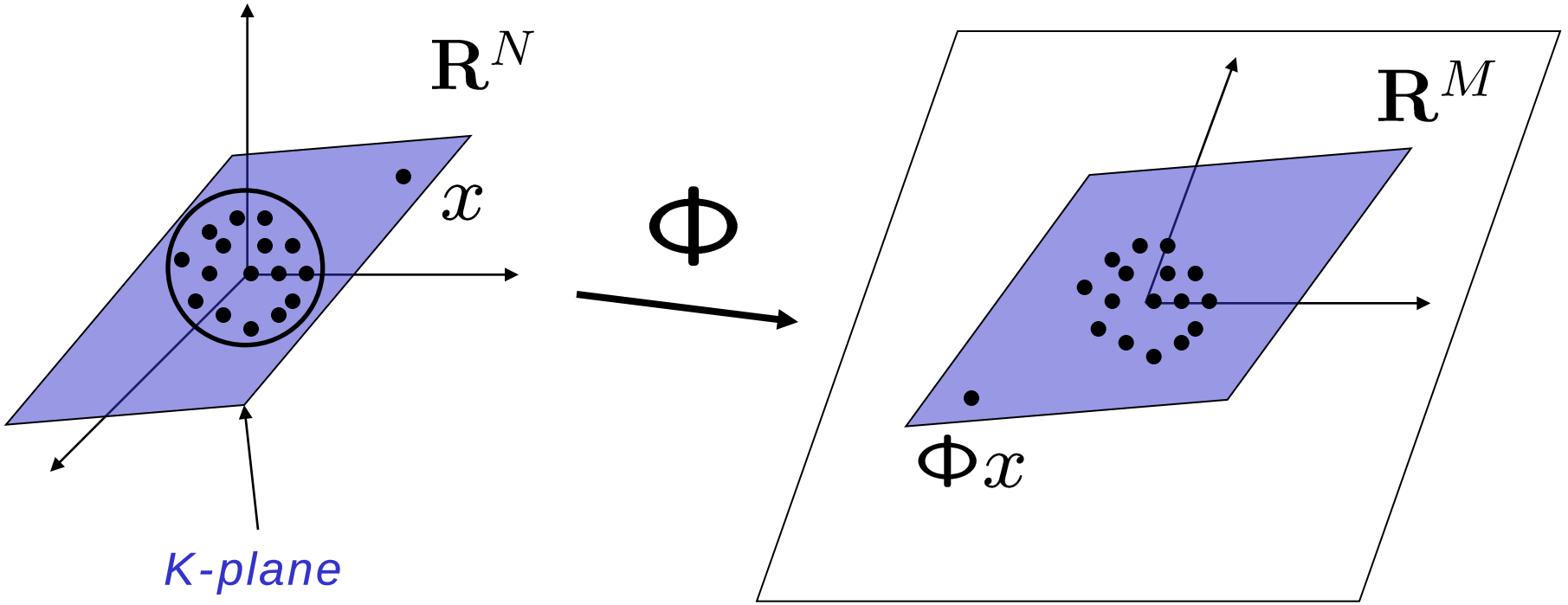
- Same techniques link JLL to RIP

[Baraniuk, Davenport, DeVore, Wakin, *Constructive Approximation*, 2008]

Connecting JL to RIP

Consider effect of random JL Φ on each K-plane

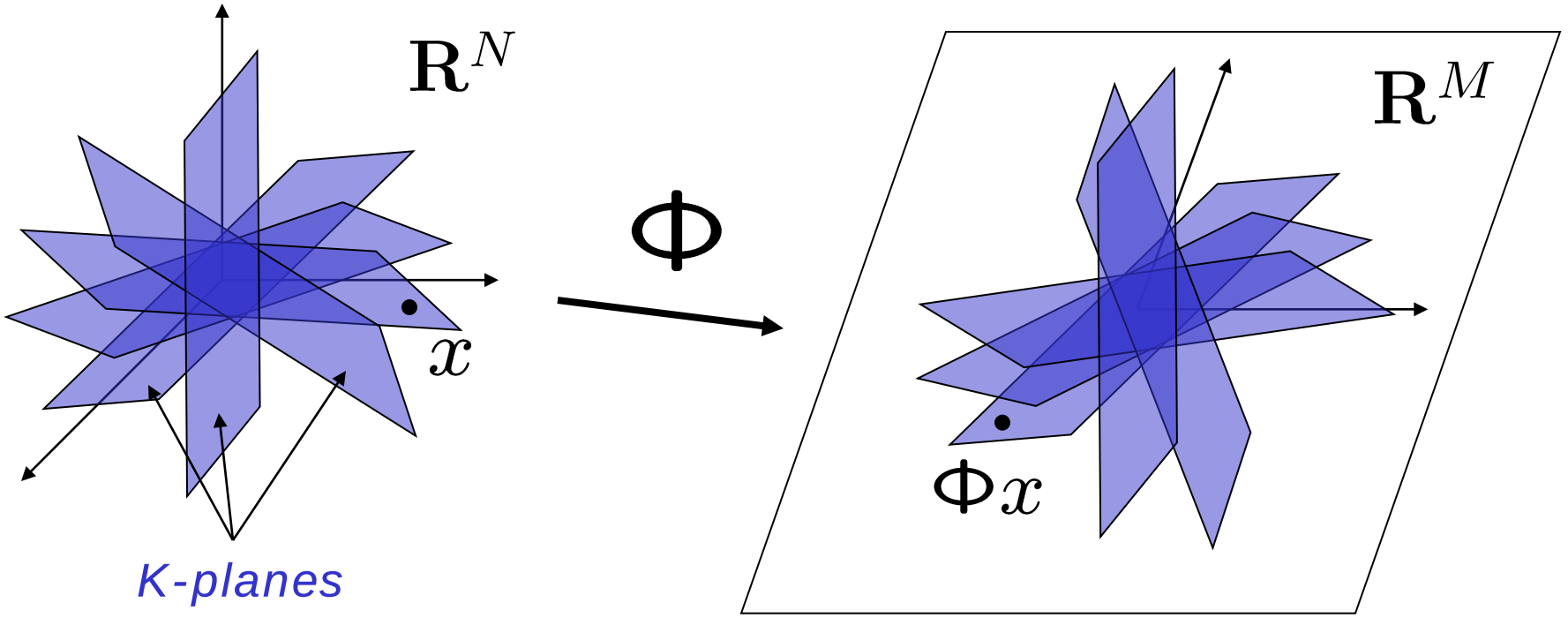
- construct covering of points Q on unit sphere
- JL: isometry for each point with high probability
- union bound ➔ isometry for all points q in Q
- extend to isometry for all x in K-plane



Connecting JL to RIP

Consider effect of random JL Φ on each K-plane

- construct covering of points Q on unit sphere
- JL: isometry for each point with high probability
- union bound ➔ isometry for all points q in Q
- extend to isometry for all x in K-plane
- union bound ➔ isometry for all K-planes



Favorable JL Distributions

- Gaussian

$$\phi_{i,j} \sim \mathcal{N}\left(0, \frac{1}{M}\right)$$

- Bernoulli/Rademacher [Achlioptas]

$$\phi_{i,j} := \begin{cases} +\frac{1}{\sqrt{M}} & \text{with probability } \frac{1}{2}, \\ -\frac{1}{\sqrt{M}} & \text{with probability } \frac{1}{2} \end{cases}$$

- "Database-friendly" [Achlioptas]

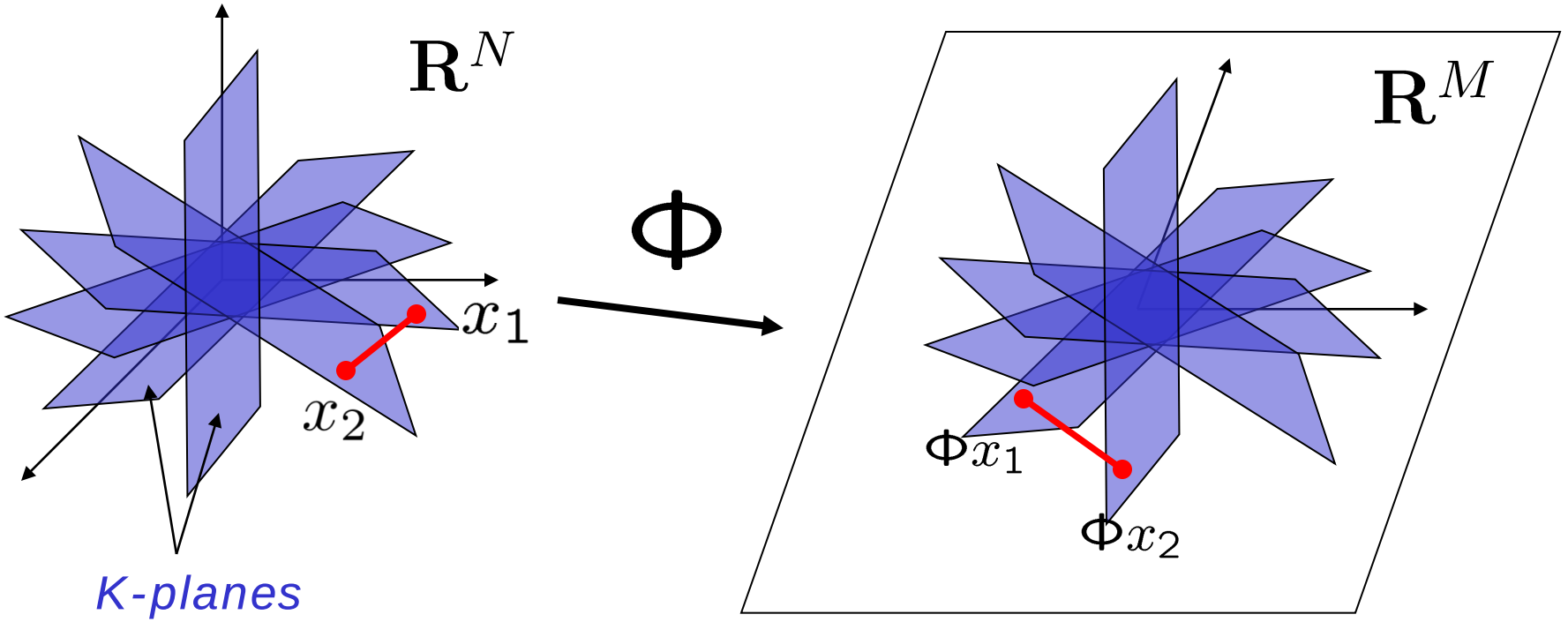
$$\phi_{i,j} := \begin{cases} +\sqrt{\frac{3}{M}} & \text{with probability } \frac{1}{6}, \\ 0 & \text{with probability } \frac{2}{3}, \\ -\sqrt{\frac{3}{M}} & \text{with probability } \frac{1}{6} \end{cases}$$

- Random Orthoprojection to R^M [Gupta, Dasgupta]

RIP as a "Stable" Embedding

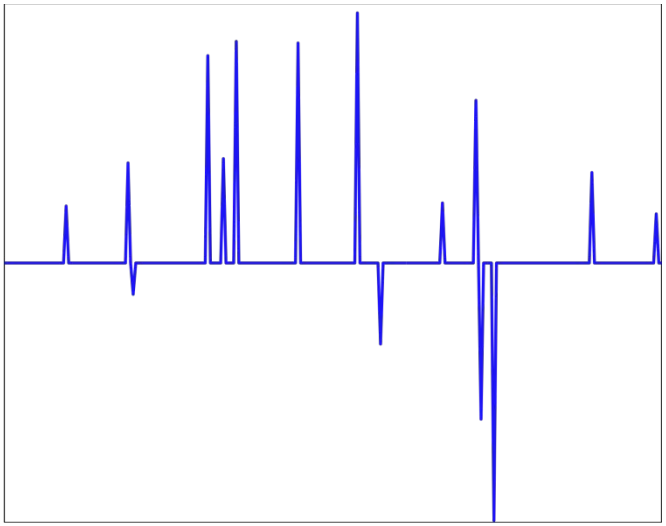
- RIP of order $2K$ implies: for all K-sparse x_1 and x_2,

$$(1-\delta_{2K}) \leq \frac{\|\Phi x_1 - \Phi x_2\|_2^2}{\|x_1 - x_2\|_2^2} \leq (1+\delta_{2K})$$

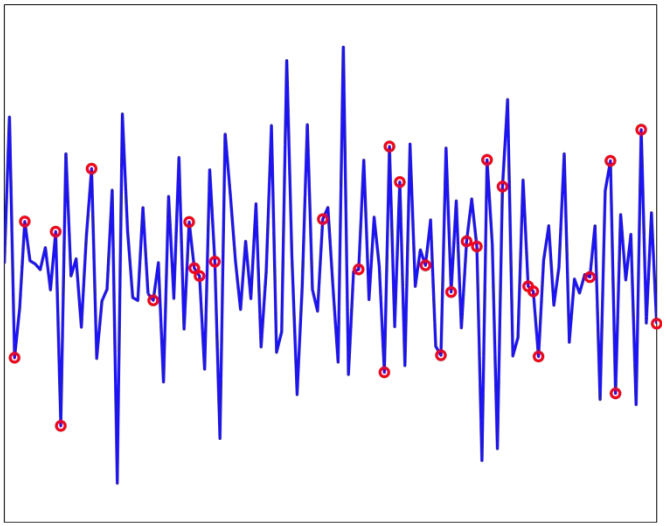


Structured Random Matrices

- There are more structured (but still random) compressed sensing matrices
- We can randomly sample in a domain whose basis vectors are *incoherent* with the sparsity basis
- Example: sparse in time, sample in frequency



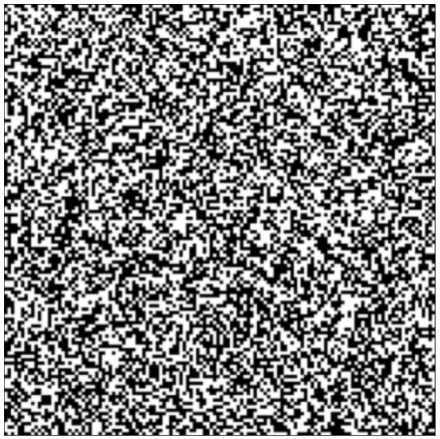
time



frequency

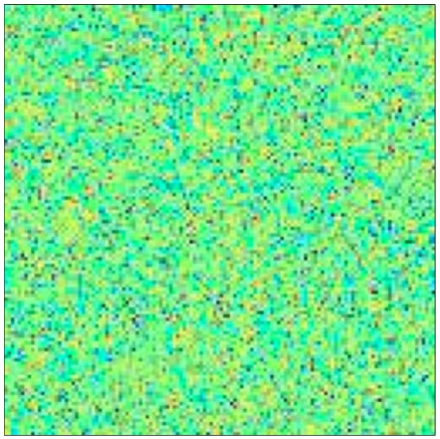
Structured Random Matrices

- Signal is sparse in the wavelet domain, measured with *noiselets* (Coifman et al. '01)



2D noiselet

wavelet domain



noiselet domain

- Stable recovery from

$$M = O(K \log^4 N)$$

measurements

Compressive Sensing

Recovery Algorithms

Sparse Recovery Algorithms

- **Goal:** given $u = \Phi x + n$

  recover x

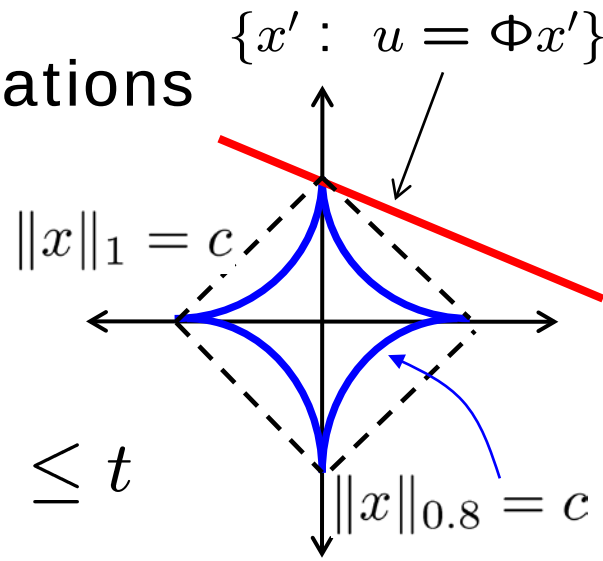
- $\ell_{q:q\leq 1}$ and convex optimization formulations
  - basis pursuit, Lasso, BP denoising...

$$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } u = \Phi x$$

$$\widehat{x} = \arg\min \|u - \Phi x\|_2 \text{ s.t. } \|x\|_1 \leq t$$

$$\widehat{x} = \arg\min \|u - \Phi x\|_2^2 + \mu \|x\|_1$$

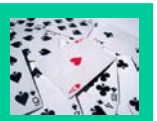
  - iterative re-weighted $\ell_1 \& \ell_2$ algorithms



- Hard thresholding algorithms: ALPS, CoSaMP, SP,...
- Greedy algorithms: OMP, MP,...

http://lions.epfl.ch/ALPS

Sparse Recovery Algorithms

	Combinatorial $\binom{N}{K}$	Geometric 	Probabilistic 
Encoding	non-convex union-of-subspaces	atomic norm / convex relaxation	compressible / sparse priors
Example	$\min_{x:\|x\|_0 \leq K} \|u - \Phi x\|^2$	$\min_{x:\|x\|_1 \leq \lambda} \|u - \Phi x\|^2$	$E\{x\|u\}$
Algorithm	IHT, CoSaMP, SP, ALPS, OMP...	Basis pursuit, Lasso, basis pursuit denoising...	Variational Bayes, EP, Approximate message passing (AMP)...

$$\|x\|_0 = \#\{x_i \neq 0\}$$

http://lions.epfl.ch/ALPS

Sparse Recovery Algorithms

The Clash Operator

	Combinatorial $\binom{N}{K}$	Geometric 	Probabilistic 
Encoding	non-convex union-of-subspaces	atomic norm / convex relaxation	compressible / sparse priors
Example	$\min_{x:\|x\|_0 \leq K} \|u - \Phi x\|^2$	$\min_{x:\|x\|_1 \leq \lambda} \|u - \Phi x\|^2$	$E\{x\|u\}$
Algorithm	IHT, CoSaMP, SP, ALPS, OMP...	Basis pursuit, Lasso, basis pursuit denoising...	Variational Bayes, EP, Approximate message passing (AMP)...

$$\widehat{x}_{\text{Clash}} = \arg\min_{x:\|x\|_0 \leq K, \|x\|_1 \leq \lambda} \|u - \Phi x\|^2$$

$$\|x\|_0 = \#\{x_i \neq 0\}$$

http://lions.epfl.ch/CLASH

L1 with equality constraints = linear programming

The standard L1 recovery program

$$\min_x \ \|x\|_{\ell_1} \quad \text{s.t.} \quad y = \Phi x$$

is equivalent to the linear program

$$\min_{x,t} \sum_i t_i \quad \text{s.t.} \quad -t_i \leq x_i \leq t_i, \quad \Phi x = y$$

There has been a tremendous amount of progress in solving linear programs in the last 15 years

SOCP

- Standard LP recovery

$$\min \|x\|_1 \quad \text{subject to } y = \Phi x$$

- Noisy measurements

$$y = \Phi x + n$$

- Second-Order Cone Program

$$\min \|x\|_1 \quad \text{subject to } \|y - \Phi x\|_2 \leq \epsilon$$

- Convex, **quadratic program**

Other Flavors of L1

- Quadratic relaxation (called LASSO in statistics)

$$\min_x \ \|x\|_{\ell_1} \ + \ \lambda \|y - \Phi x\|_2^2$$

- Dantzig selector (residual correlation constraints)

$$\min_x \ \|x\|_{\ell_1} \quad \text{s.t.} \quad \|\Phi^T (y - \Phi x)\|_\infty$$

- L1 Analysis (Ψ is an overcomplete frame)

$$\min_x \ \|\Psi^T x\|_{\ell_1} \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \epsilon$$

Stability

- Recovery is robust against noise and modeling error
- Suppose we observe

$$y = \Phi x_0 + e, \quad \|e\|_2 \leq \epsilon$$

- Relax the recovery algorithm, solve

$$\min_x \|x\|_{\ell_1} \quad \text{subject to} \quad \|y - \Phi x\|_2 \leq \epsilon$$

- The recovery error obeys

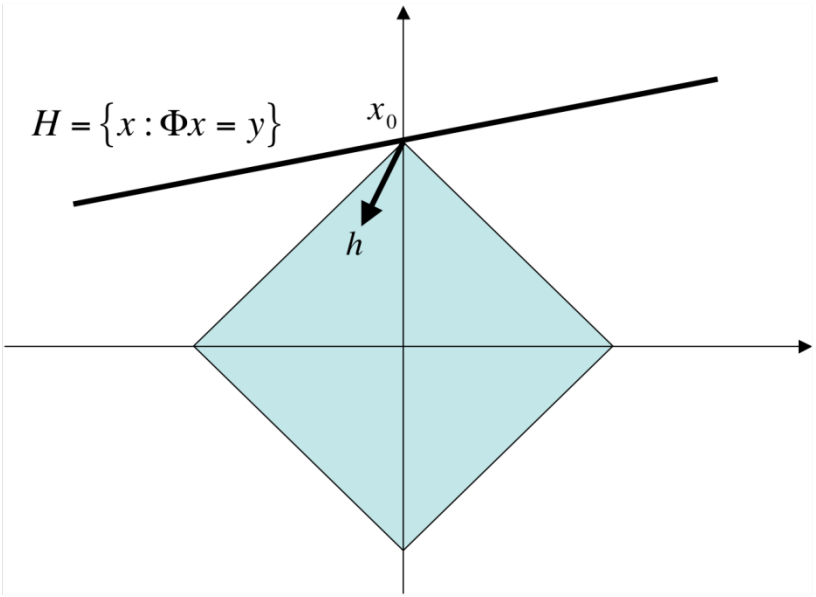
$$\|x^* - x_0\|_2 \lesssim \epsilon + \frac{\|x_{0,K} - x_0\|_{\ell_1}}{\sqrt{K}}$$

measurement error + *approximation error*

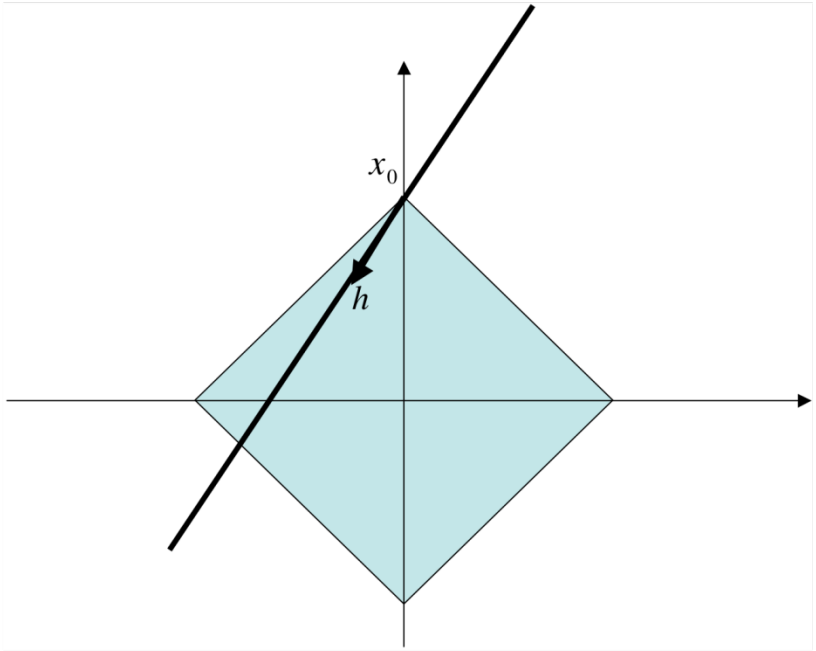
$x_{0,K} =$ best K-term approximation

Geometrical Viewpoint, Noiseless

good



bad



- Consider and "ℓ_1-descent vectors" h for feasible x:

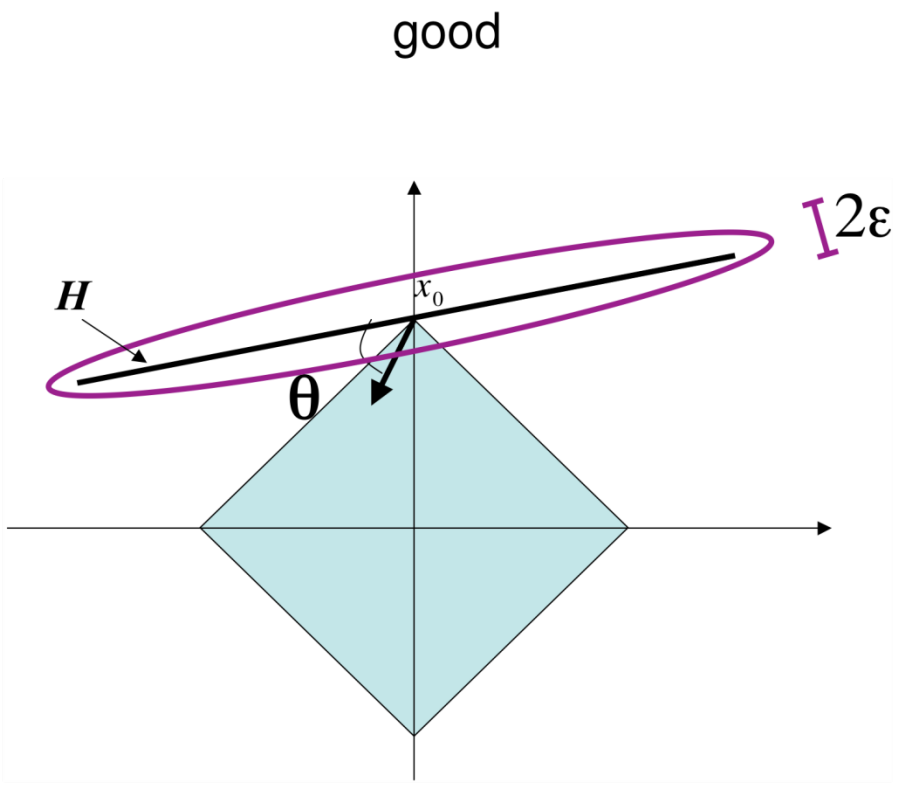
$$\|x_0 + h\|_{\ell_1} < \|x_0\|_{\ell_1}$$

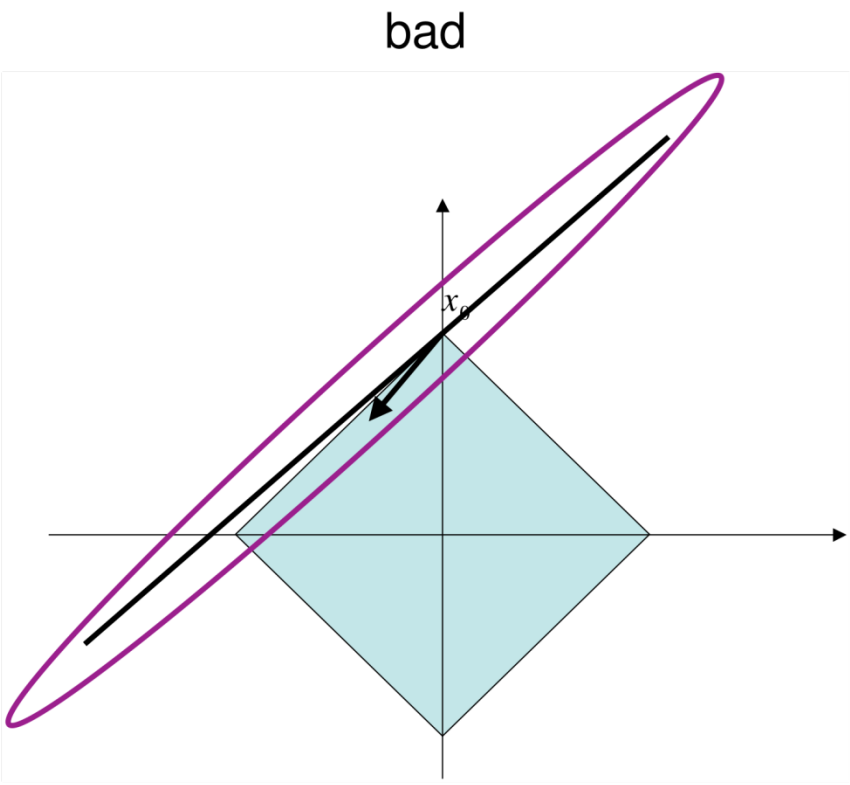
- x_0 is the solution if

$$\Phi h \neq 0$$

  for all such descent vectors

Geometrical Viewpoint, Noise





- Solution will be within ϵ of $\boldsymbol{H}$
- Need that not too much of the ℓ_1 ball near x_0 is feasible

Solving L1

- "Classical" (mid-90s) interior point methods
  - main building blocks due to Nemirovski
  - second-order, series of local quadratic approximations
  - boils down to a series of linear systems of equations
  - formulation is very general (and hence adaptable)
- Modern progress (last 5 years) has been on "first order" methods
  - Main building blocks due to Nesterov (mid 80s)
  - iterative, require applications of Φ and Φ^T at each iteration
  - convergence in 10s-100s of iterations typically
- Many software packages available
  - Fixed-point continuation (Rice)
  - Bregman iteration-based methods (UCLA)
  - NESTA (Caltech)
  - GPSR (Wisconsin)
  - SPGL1 (UBC)…..

Probabilistic View

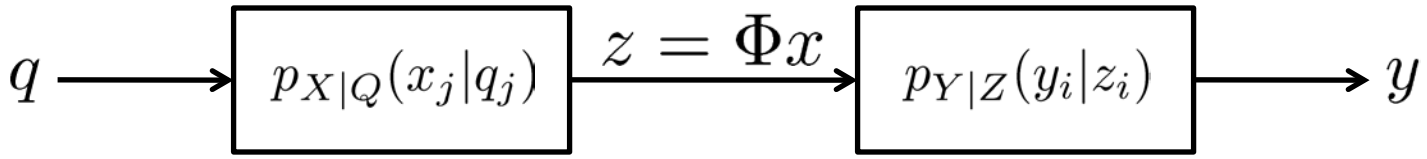
An Introduction to Approximate Message Passing (AMP)

Probabilistic Sparse Recovery: the AMP Way

- **Goal:** given $y = \Phi x + w$

  infer $E\{x|y\}$ and $\mathrm{var}\{x|y\}$

- **Model:**



Example: $p_{X|Q}(x|q = [\lambda, \hat{\theta}, \mu^\theta]) = \lambda \mathcal{N}(x; \hat{\theta}, \mu^\theta) + (1-\lambda)\delta(x), \quad \lambda \in [0,1]$

$$p_{Y|Z}(y|z) = \mathcal{N}(y; z, \mu^w)$$

- **Approach:** graphical models / message passing
- **Key Ideas:** **CLT / Gaussian approximations**

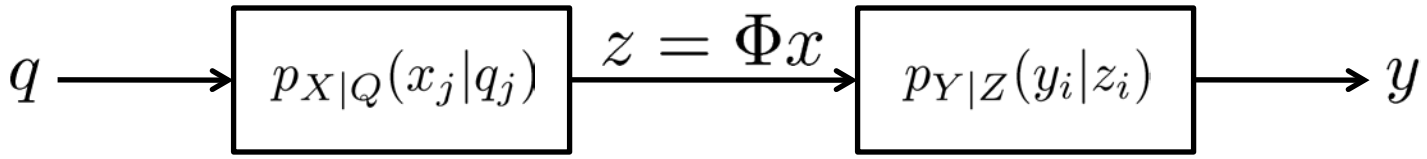
[Boutros and Caire 2002; Montanari and Tse 2006; Guo and Wang 2006; Tanaka and Okada 2006; Donoho, Maleki, and Montanari 2009; Rangan 2010]

Probabilistic Sparse Recovery: the AMP Way

- **Goal:** given $y = \Phi x + w$

  infer $E\{x|y\}$ and $\mathrm{var}\{x|y\}$

- **Model:**



Example: $p_{X|Q}(x|q = [\lambda, \hat{\theta}, \mu^\theta]) = \lambda \mathcal{N}(x; \hat{\theta}, \mu^\theta) + (1-\lambda)\delta(x), \quad \lambda \in [0,1]$

$$p_{Y|Z}(y|z) = \mathcal{N}(y; z, \mu^w)$$

- **Approach:** graphical models / message passing
- **Key Ideas:** **CLT / Gaussian approximations**

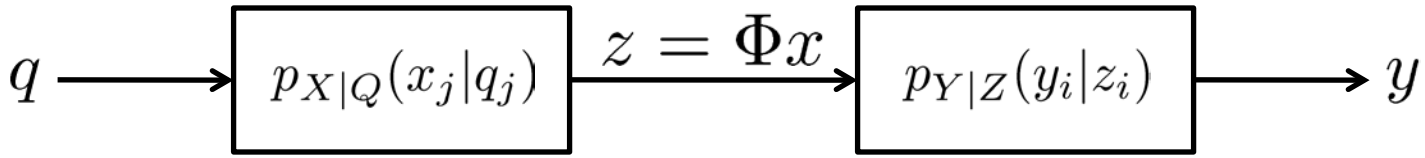
[Boutros and Caire 2002; Montanari and Tse 2006; Guo and Wang 2006; Tanaka and Okada 2006; Donoho, Maleki, and Montanari 2009; Rangan 2010]

Probabilistic Sparse Recovery: the AMP Way

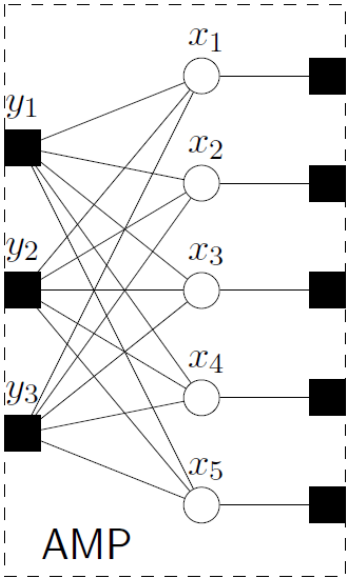
- **Goal:** given $y = \Phi x + w$

  infer $E\{x|y\}$ and $\mathrm{var}\{x|y\}$

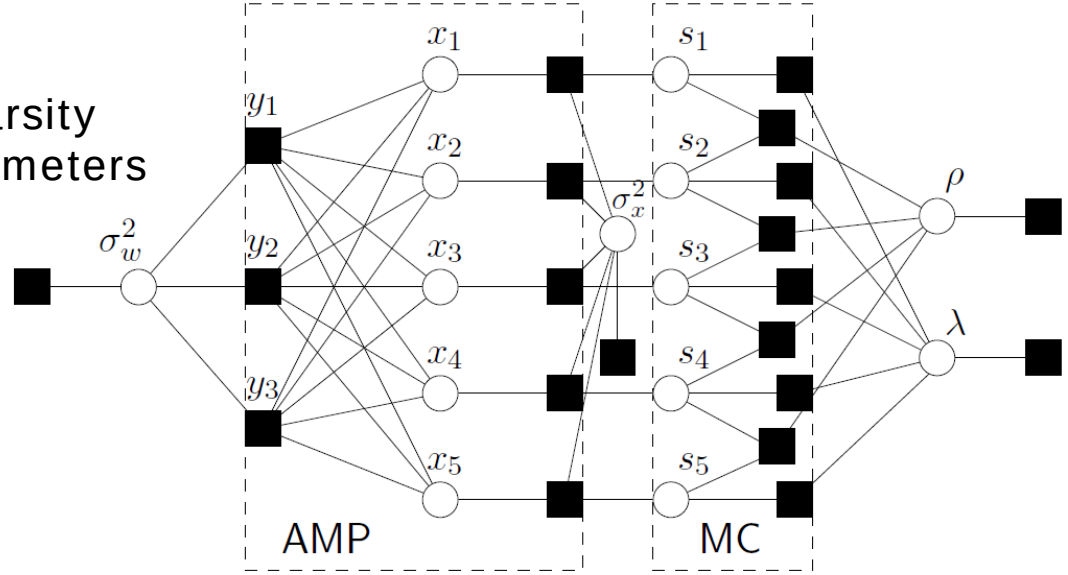
- **Model:**



- **Approach:** graphical models are modular!



structured sparsity
unknown parameters



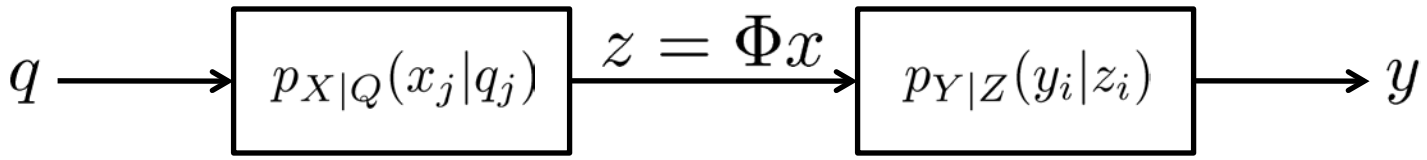
Probabilistic Sparse Recovery: the AMP Way

- **Goal:** given $y = \Phi x + w$

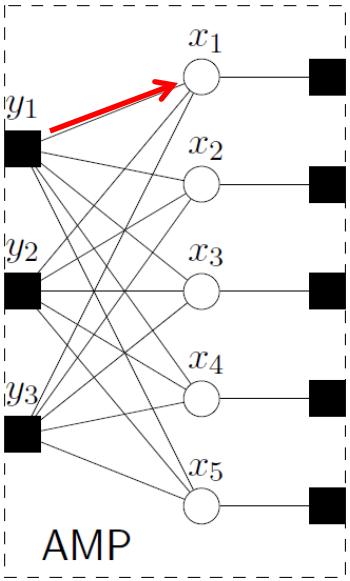
  infer $E\{x|y\}$ and $\mathrm{var}\{x|y\}$

- **Model:**

$$p_{Y|Z}(y|z) = \mathcal{N}(y; z, \mu^w)$$



- **Approximate message passing:** (sum-product)



$$m_{i\to j}(x_j) = E_{x_r:r\neq j}\{p_{Y|Z}(y_i|z_i)|x_j, y_i\}$$

$$= E_{x_r:r\neq j}\left\{p_{Y|Z}\left(y_i\left|[\Phi]_{ij}x_j + \sum_{r\neq j}[\Phi]_{ir}x_r\right.\right)|x_j, y_i\right\}$$

$$\approx E_V\left\{p_{Y|Z}\left(y_i\,|[\Phi]_{ij}x_j + V\right)|x_j, y_i\right\}$$

$$\approx d_1[\Phi]_{ij}x_j + d_2[\Phi]_{ij}x_j^2$$

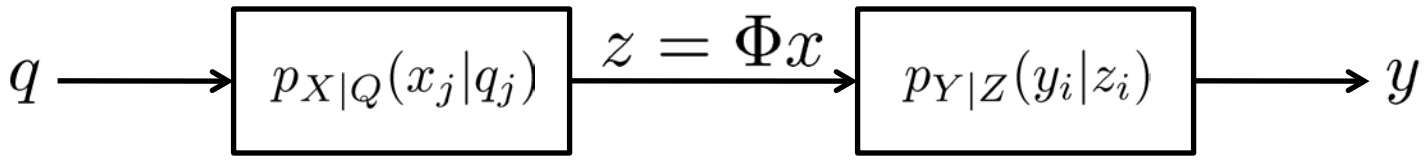
Probabilistic Sparse Recovery: the AMP Way

- **Goal:** given $y = \Phi x + w$

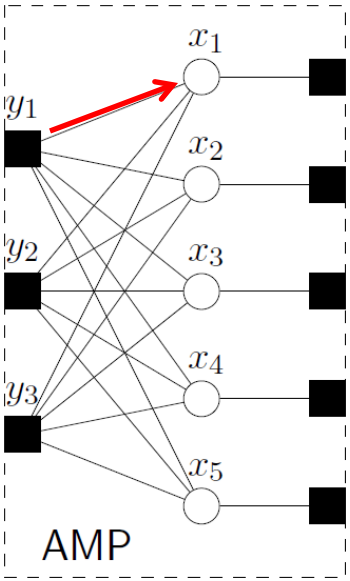
  infer $E\{x|y\}$ and $\mathrm{var}\{x|y\}$

- **Model:**

$$p_{Y|Z}(y|z) = \mathcal{N}(y; z, \mu^w)$$



- **Approximate message passing:** (sum-product)



$$m_{i\to j}(x_j) = E_{x_r: r\neq j}\{p_{Y|Z}(y_i|z_i)|x_j, y_i\}$$

$$= E_{x_r: r\neq j}\left\{p_{Y|Z}\left(y_i \middle| [\Phi]_{ij}x_j + \sum_{r\neq j}[\Phi]_{ir}x_r\right)|x_j, y_i\right\}$$

$$\approx \mathbf{E_V}\left\{\mathbf{p_{Y|Z}}\left(\mathbf{y_i}\,|[\mathbf{\Phi}]_{\mathbf{ij}}\mathbf{x_j} + \mathbf{V}\right)|\mathbf{x_j}, \mathbf{y_i}\right\} \qquad V \sim \mathcal{N}$$

$$\approx d_1[\Phi]_{ij}x_j + d_2[\Phi]_{ij}x_j^2$$

Central limit theorem (blessing-of-dimensionality)

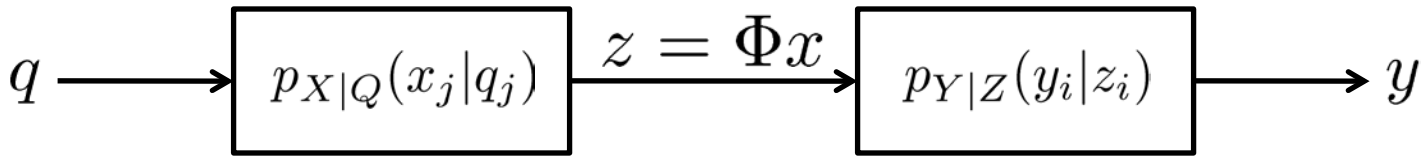
Probabilistic Sparse Recovery: the AMP Way

- **Goal:** given $y = \Phi x + w$

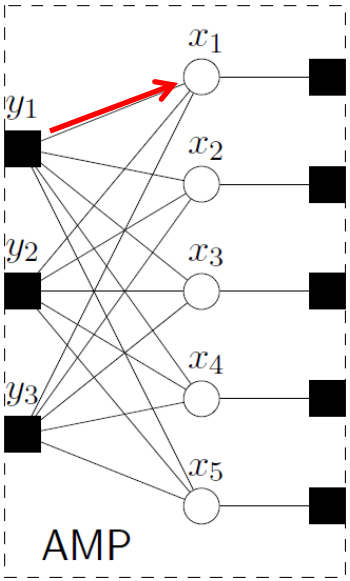
  infer $E\{x|y\}$ and $\mathrm{var}\{x|y\}$

- **Model:**

$$p_{Y|Z}(y|z) = \mathcal{N}(y; z, \mu^w)$$



- **Approximate message passing:** (sum-product)



$$m_{i\to j}(x_j) = E_{x_r:r\neq j}\{p_{Y|Z}(y_i|z_i)|x_j, y_i\}$$

$$= E_{x_r:r\neq j}\left\{p_{Y|Z}\left(y_i \middle| [\Phi]_{ij}x_j + \sum_{r\neq j}[\Phi]_{ir}x_r\right)|x_j, y_i\right\}$$

$$\approx E_V\left\{p_{Y|Z}\left(y_i\,|[\Phi]_{ij}x_j + V\right)|x_j, y_i\right\}$$

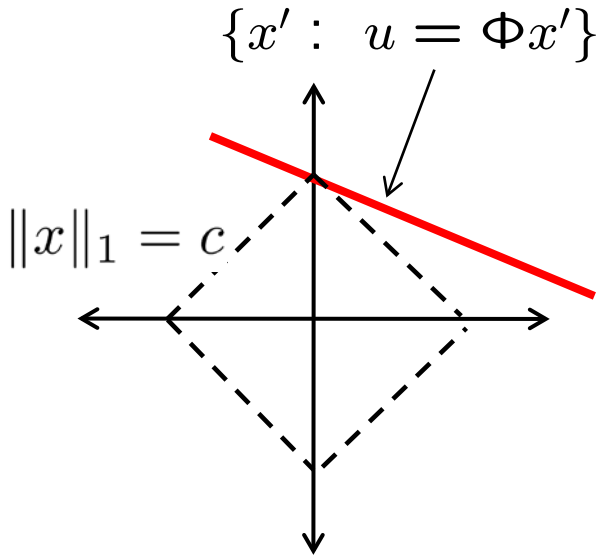
$$\approx \mathbf{d_1[\Phi]_{ij}x_j + d_2[\Phi]_{ij}x_j^2}$$

Taylor series approximation

Importance of Geometry

$\widehat{x} = \arg\min \|x\|_0 \text{ s.t. } u = \Phi x$

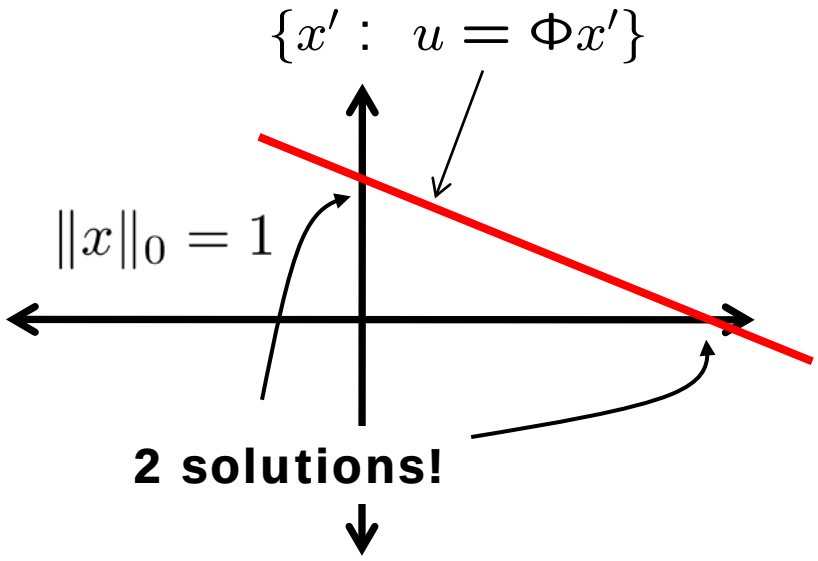
$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } u = \Phi x$



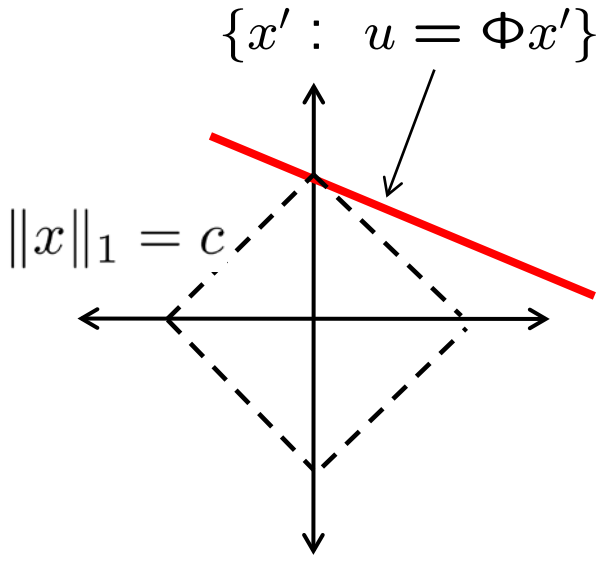
Importance of Geometry

- A subtle issue

$\widehat{x} = \arg\min \|x\|_0 \text{ s.t. } u = \Phi x$



$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } u = \Phi x$

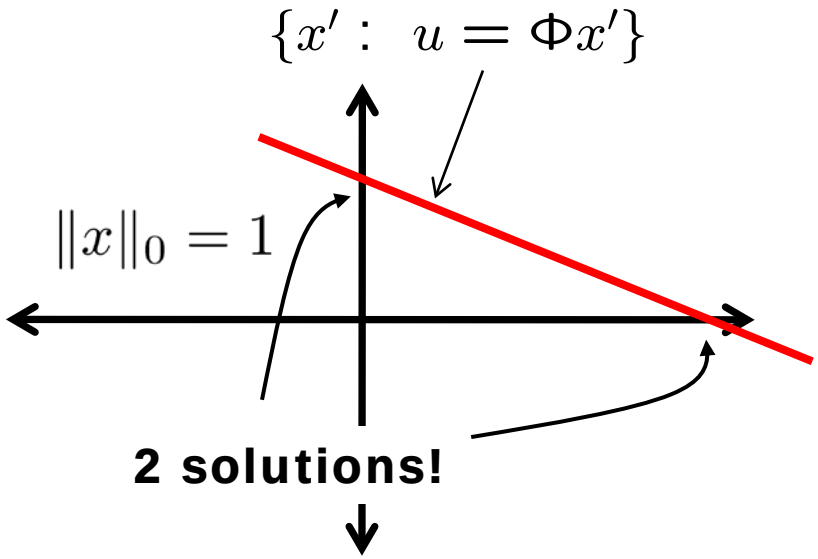


Which one is correct?

Importance of Geometry

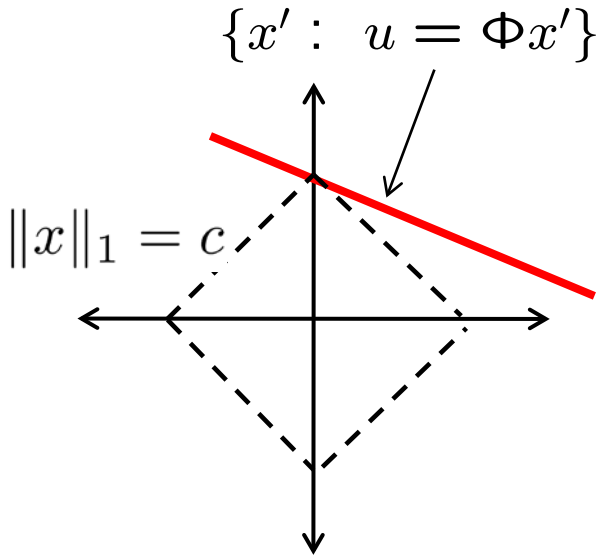
- A subtle issue

$$\widehat{x} = \arg\min \|x\|_0 \text{ s.t. } u = \Phi x$$



Which one is correct?

$$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } u = \Phi x$$

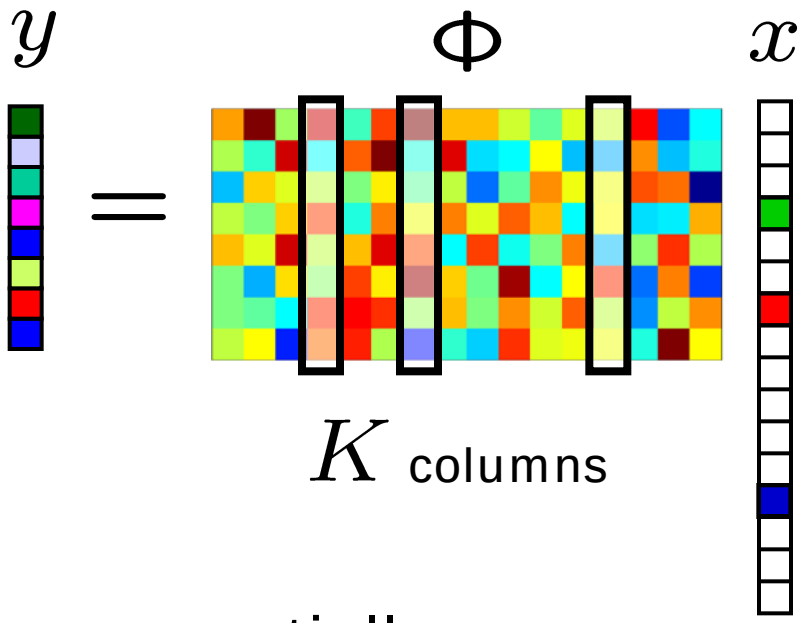


Matching Pursuit

- Greedy algorithm

- **Key ideas:**
  (1) measurements y composed of sum of *K* columns of Φ

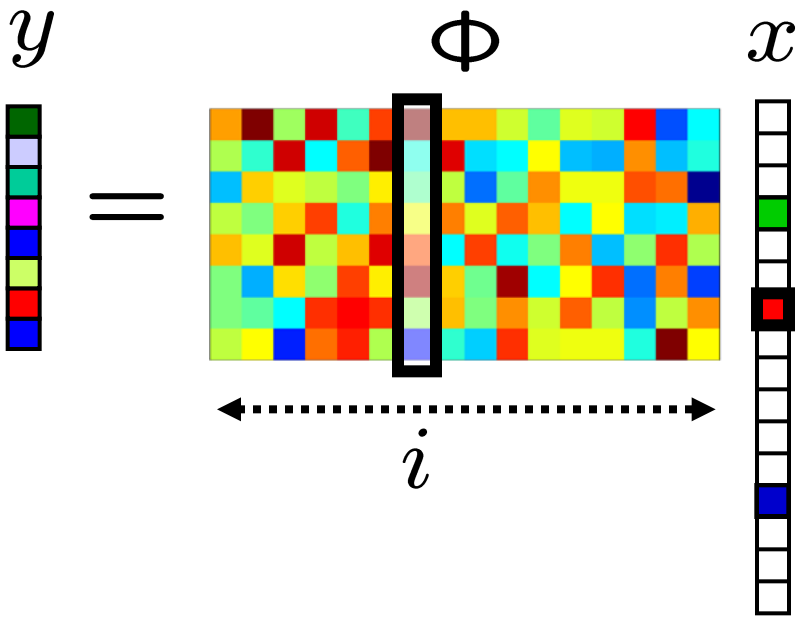
  (2) identify which *K* columns sequentially according to size of contribution to y



Matching Pursuit

- For each column ϕ_i compute

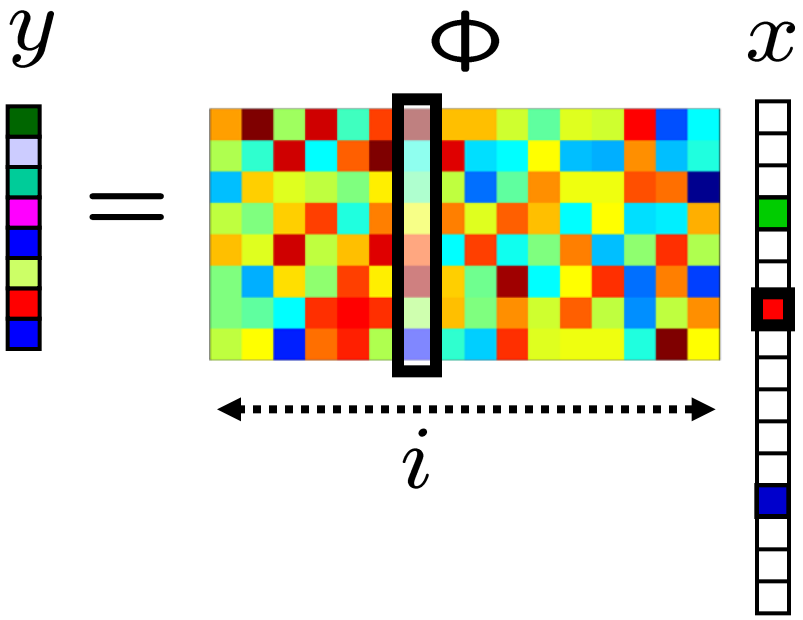
$$\widehat{x}_i = \langle y, \phi_i \rangle$$



- Choose largest $|\widehat{x}_i|$ (greedy)
- Update estimate $\widehat{x}$ by adding in $\widehat{x}_i$
- Form residual measurement $y' = y - x_i \phi_i$ and iterate until convergence

Orthogonal Matching Pursuit

- Same procedure as Matching Pursuit



- Except at each iteration:
  - remove selected column ϕ_i
  - re-orthogonalize the remaining columns of Φ
- Converges in *K* iterations

CoSAMP

- Needell and Tropp, 2008
- Very simple greedy algorithm, provably effective

ALGORITHM 2.1: CoSaMP Recovery Algorithm

CoSaMP($\mathbf{\Phi}$, $\boldsymbol{u}$, s)
Input: Sampling matrix $\mathbf{\Phi}$, noisy sample vector $\boldsymbol{u}$, sparsity level s
Output: An s-sparse approximation $\boldsymbol{a}$ of the target signal

---

$\boldsymbol{a}^0 \leftarrow \mathbf{0}$ { Trivial initial approximation }
$\boldsymbol{v} \leftarrow \boldsymbol{u}$ { Current samples = input samples }
$k \leftarrow 0$

repeat
  $k \leftarrow k + 1$

  $\boldsymbol{y} \leftarrow \mathbf{\Phi}^* \boldsymbol{v}$ { Form signal proxy }
  $\Omega \leftarrow \text{supp}(\boldsymbol{y}_{2s})$ { Identify large components }
  $T \leftarrow \Omega \cup \text{supp}(\boldsymbol{a}^{k-1})$ { Merge supports }

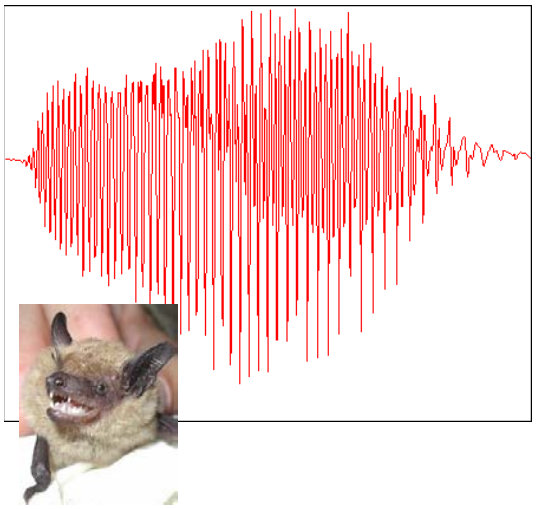
  $\boldsymbol{b}|_T \leftarrow \mathbf{\Phi}_T^{\dagger} \boldsymbol{u}$ { Signal estimation by least-squares }
  $\boldsymbol{b}|_{T^c} \leftarrow \mathbf{0}$

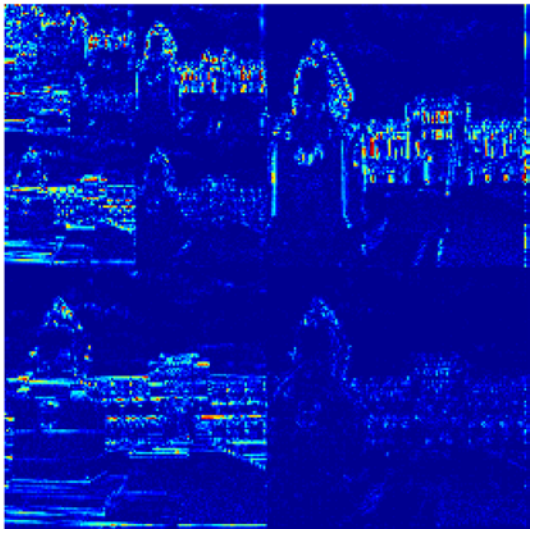
  $\boldsymbol{a}^k \leftarrow \boldsymbol{b}_s$ { Prune to obtain next approximation }
  $\boldsymbol{v} \leftarrow \boldsymbol{u} - \mathbf{\Phi} \boldsymbol{a}^k$ { Update current samples }

until halting criterion *true*

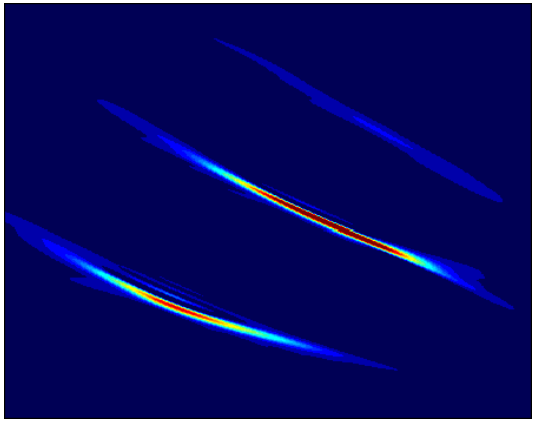
From Sparsity to Model-based (*structured*) Sparsity

Sparse Models

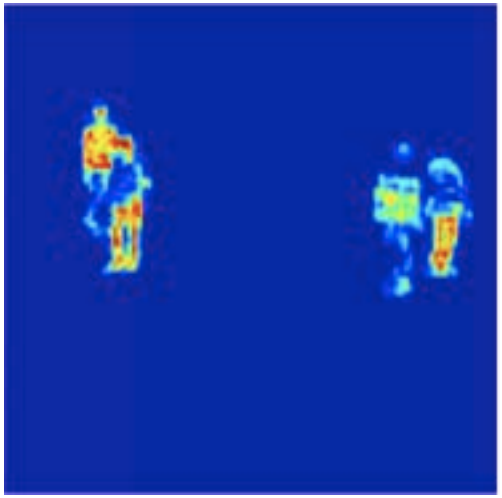




wavelets:
natural images



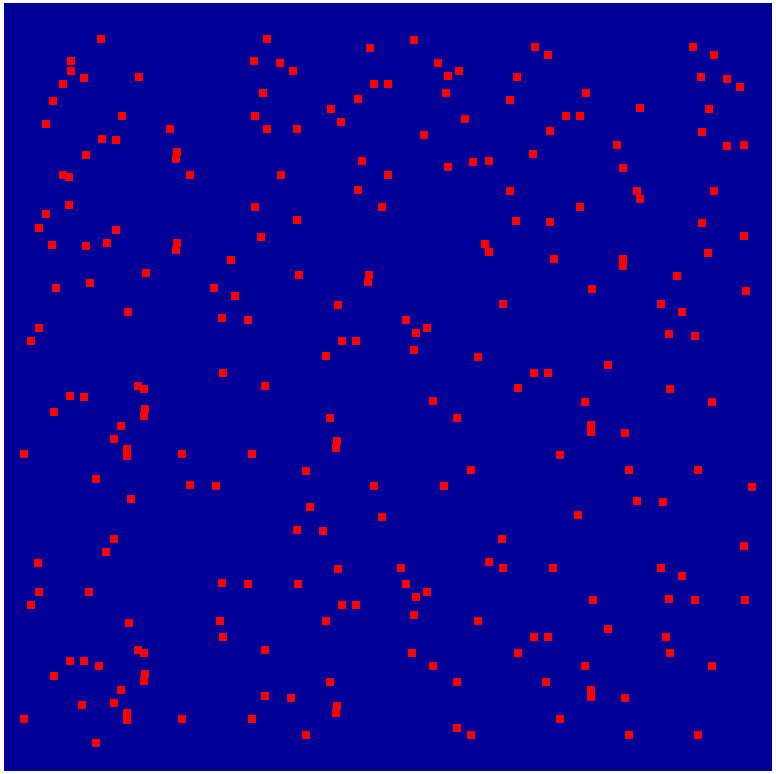
Gabor atoms:
chirps/tones



pixels:
background subtracted
images

Sparse Models

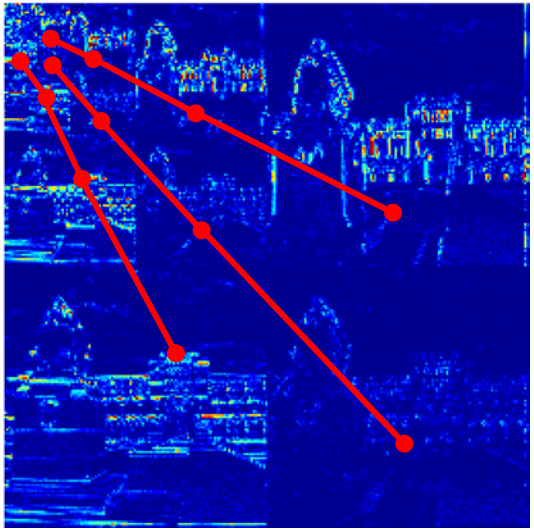
- Sparse/compressible signal model captures **simplistic primary structure**



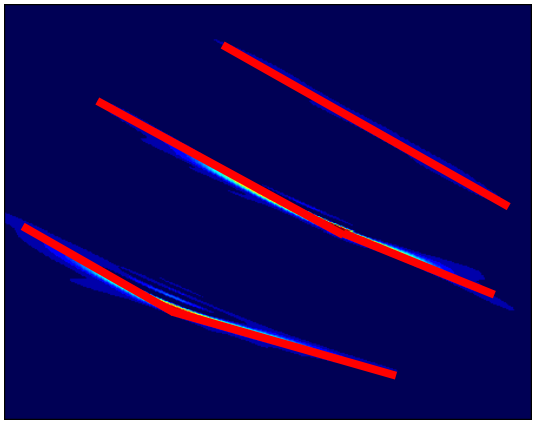
sparse image

Beyond Sparse Models

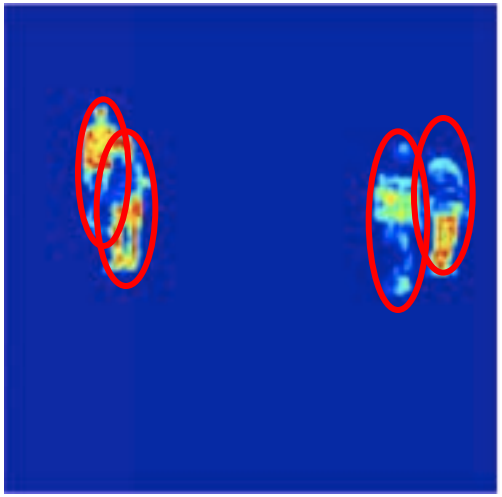
- Sparse/compressible signal model captures **simplistic primary structure**
- Modern compression/processing algorithms capture **richer secondary coefficient structure**



wavelets:
natural images



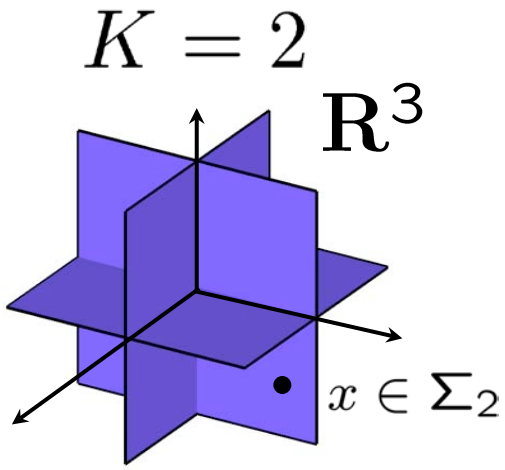
Gabor atoms:
chirps/tones

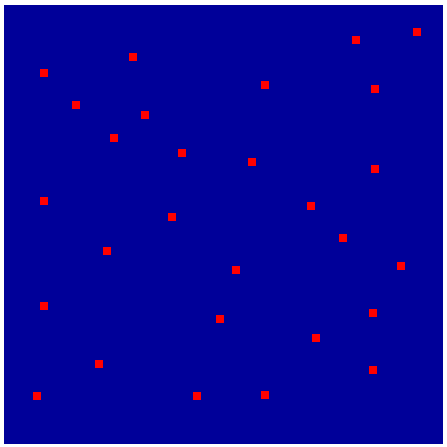


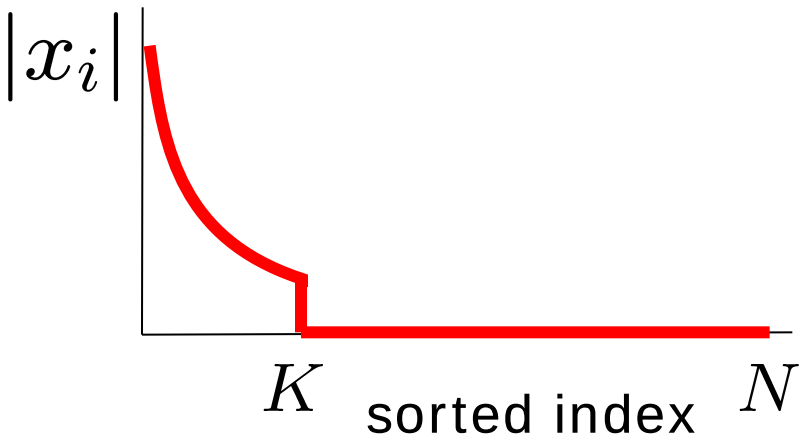
pixels:
background subtracted
images

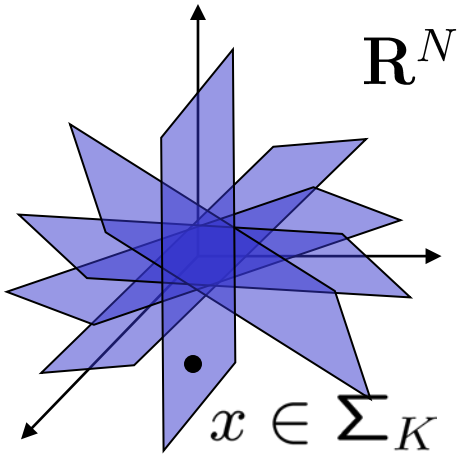
Signal Priors

- **Sparse** signal: only K out of N coordinates nonzero
  - model: union of all K-dimensional subspaces aligned w/ coordinate axes





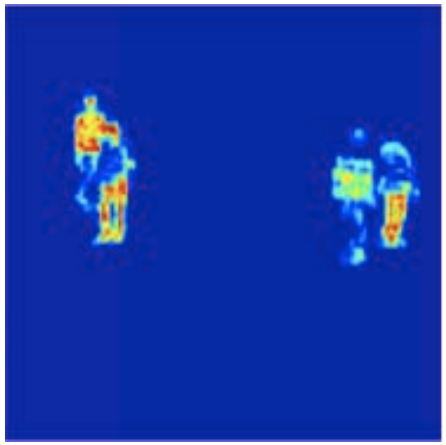


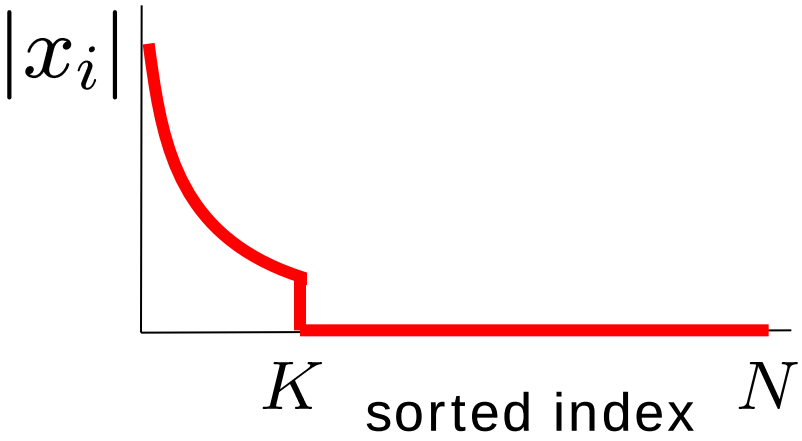


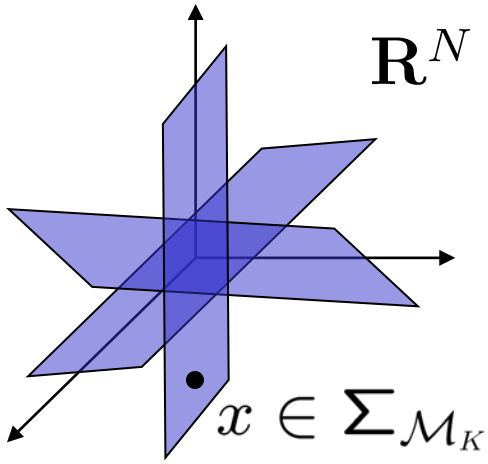
Signal Priors

- **Sparse** signal: only K out of N coordinates nonzero
  - model: union of all K-dimensional subspaces aligned w/ coordinate axes

- **Structured sparse** signal: reduced set of subspaces **(or model-sparse)**
  - model: a particular union of subspaces ex: clustered or dispersed sparse patterns







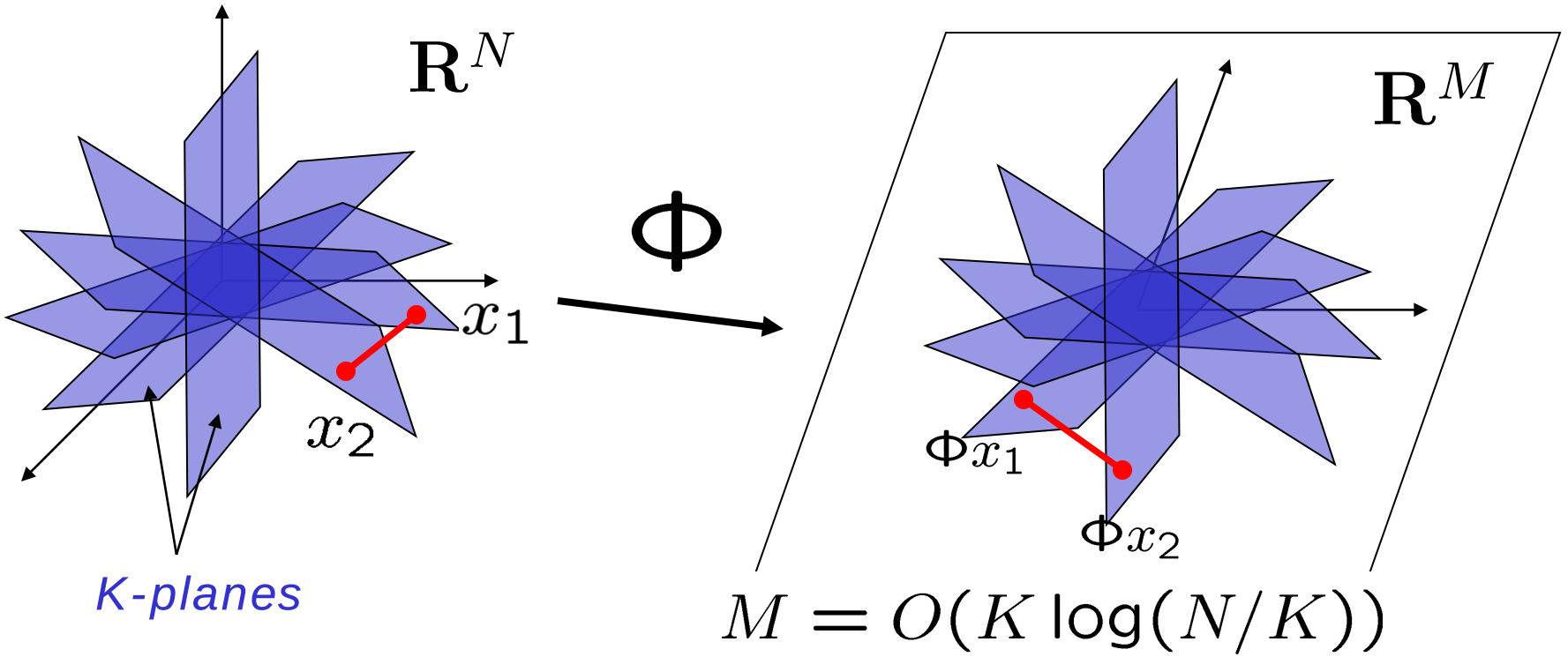
Restricted Isometry Property (RIP)

- **Model:** *K*-sparse

$$(1-\delta_{2K}) \leq \frac{\|\Phi x_1 - \Phi x_2\|_2^2}{\|x_1 - x_2\|_2^2} \leq (1+\delta_{2K})$$

- **RIP:** stable embedding

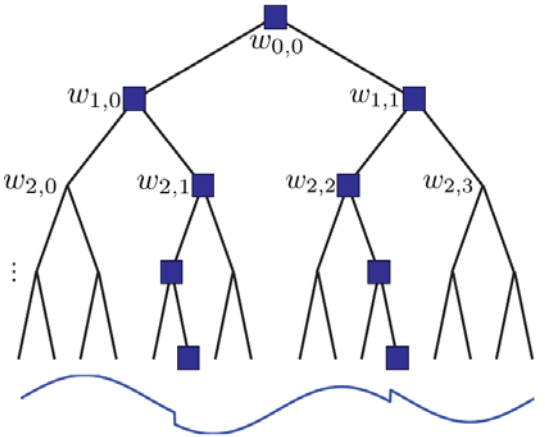
Random subGaussian (iid Gaussian, Bernoulli) matrix **< >** RIP w.h.p.

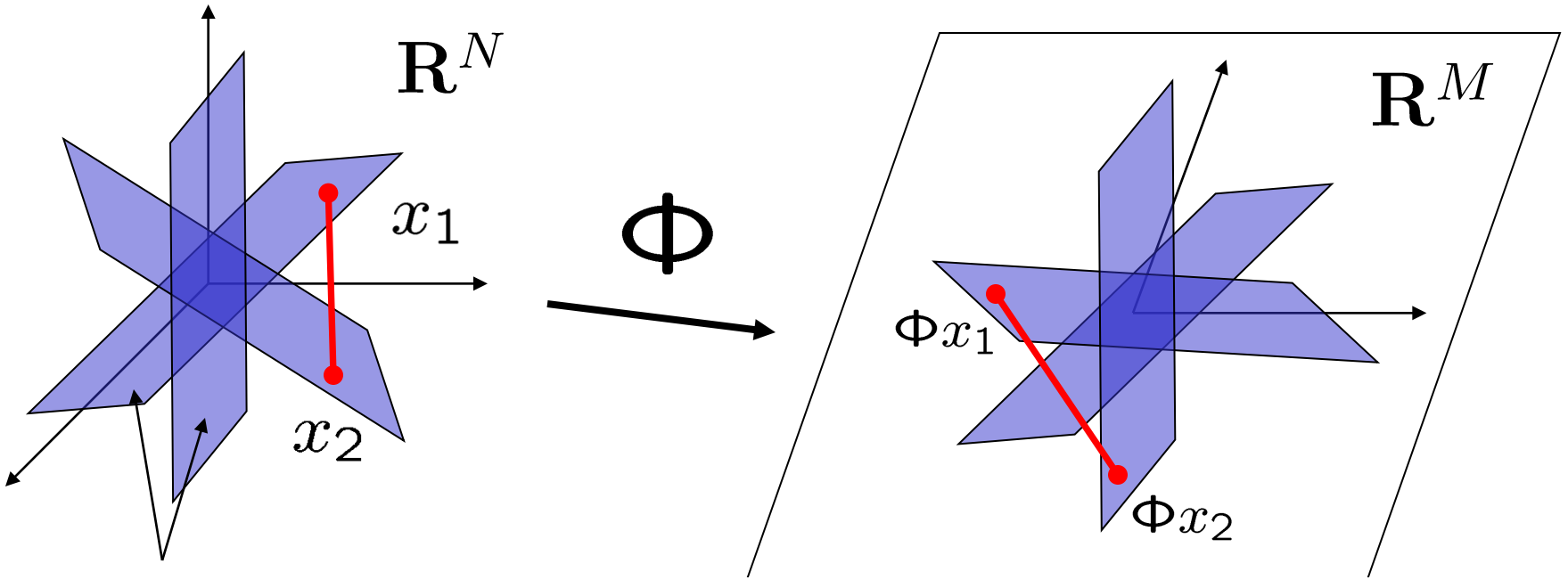


Restricted Isometry Property (RIP)

- **Model:** *K*-sparse
  + significant coefficients lie on a rooted subtree
  (a known model for piecewise smooth signals)
- **Tree-RIP:** stable embedding

$\mathcal{M}$



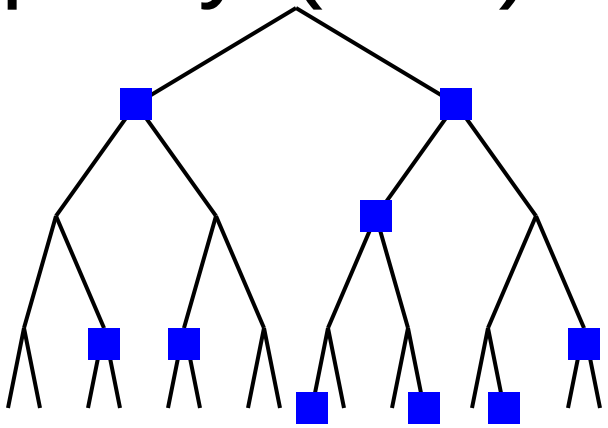


K-planes

$$\delta_{\mathcal{M}_{2K}}: \underline{M = O(K)} < O(K \log(N/K))$$

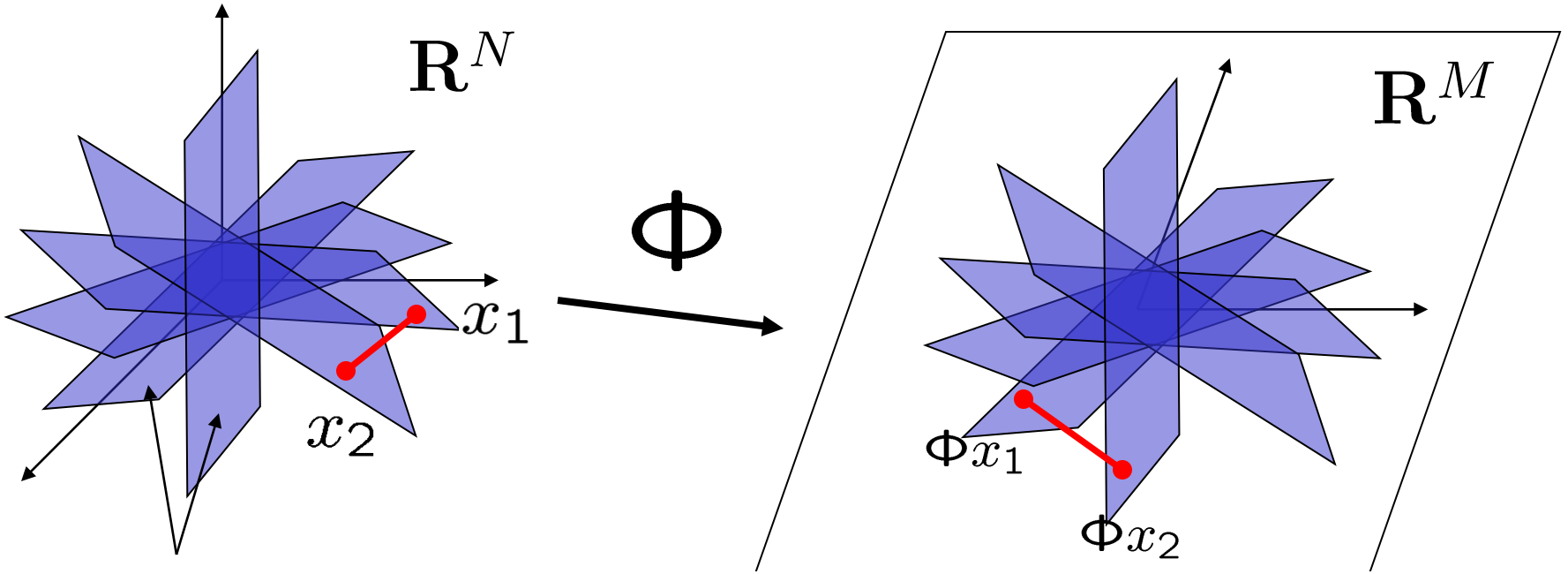
Restricted Isometry Property (RIP)

- **Model:** *K*-sparse



Note the difference:

- **RIP:** stable embedding

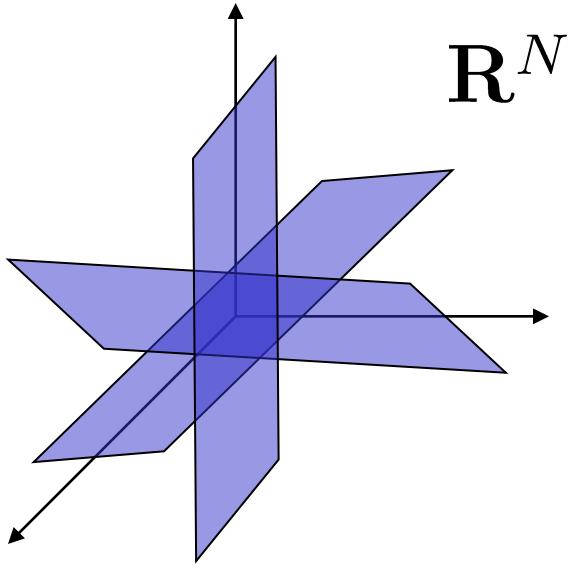


K-planes

δ_{2K}: $M = O(K \log(N/K))$

Model-Sparse Signals

- Defn: A ***K*-sparse signal *model*** comprises a particular (*reduced*) set of *K*-dim canonical subspaces

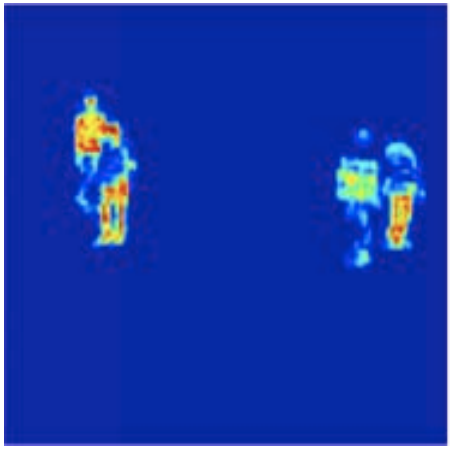


- **Structured subspaces**

  < > ***fewer measurements***

  < > ***improved recovery perf.***

  < > ***faster recovery***



CS Recovery

- **Iterative Hard Thresholding (IHT)**

  [Nowak, Figueiredo; Kingsbury, Reeves; Daubechies, Defrise, De Mol; Blumensath, Davies; ...]

  Given $y = \Phi x$, recover a sparse x

  initialize: $\widehat{x}_0 = 0,\ r = y,\ i = 0$

  iteration:

  - $i \leftarrow i + 1$
  - $b \leftarrow \widehat{x}_{i-1} + \Phi^T r$ **update signal estimate**
  - $\widehat{x}_i \leftarrow \text{thresh}(b, K)$ **prune signal estimate** (best *K*-term approx)
  - $r \leftarrow y - \Phi \widehat{x}_i$ **update residual**

  return: $\widehat{x} \leftarrow \widehat{x}_i$

Model-based CS Recovery

- **Iterative Model Thresholding**

  [**VC**, Duarte, Hegde, Baraniuk; Baraniuk, **VC**, Duarte, Hegde]

  Given $y = \Phi x$, recover a model sparse $x \in \mathcal{M}$

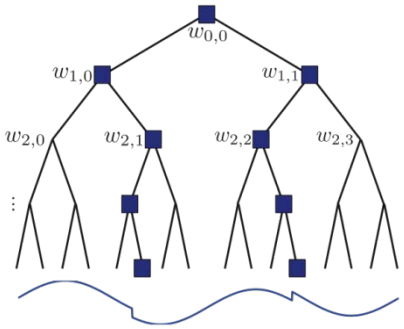
  initialize: $\widehat{x}_0 = 0,\ r = y,\ i = 0$

  iteration:

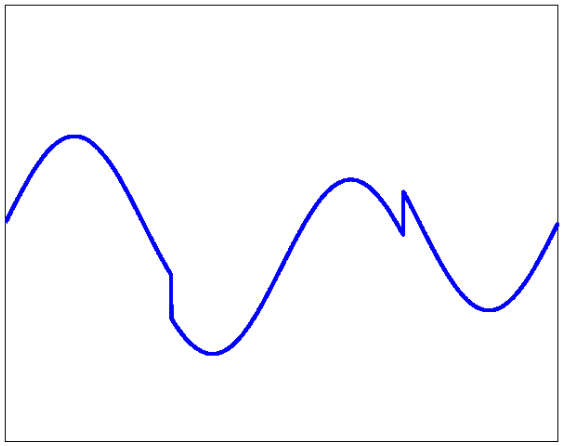
  - $i \leftarrow i + 1$
  - $b \leftarrow \widehat{x}_{i-1} + \Phi^T r$ **update signal estimate**
  - $\widehat{x}_i \leftarrow \mathcal{M}(b, K)$ **prune signal estimate** (best *K*-term **model** approx)
  - $r \leftarrow y - \Phi \widehat{x}_i$ **update residual**

  return: $\widehat{x} \leftarrow \widehat{x}_i$

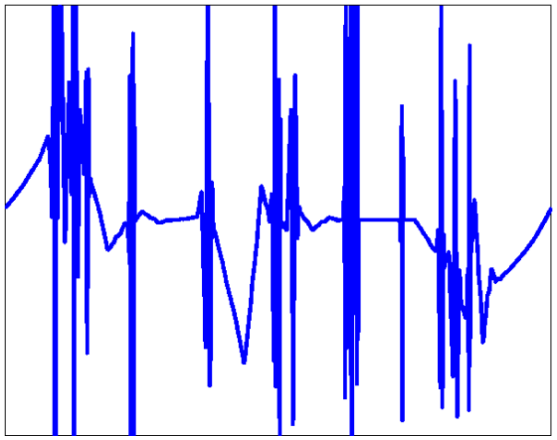
Tree-Sparse Signal Recovery



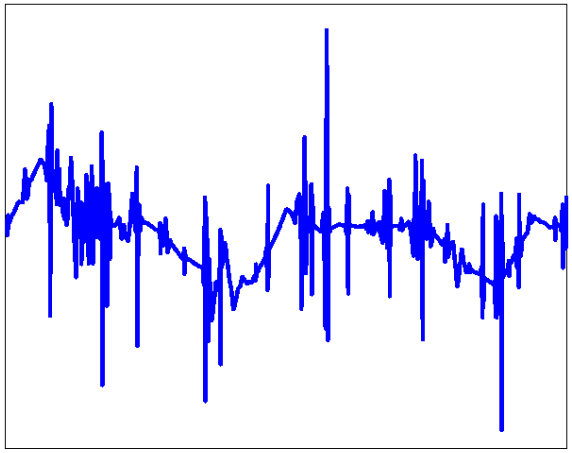
$N = 1024$
$M = 80$



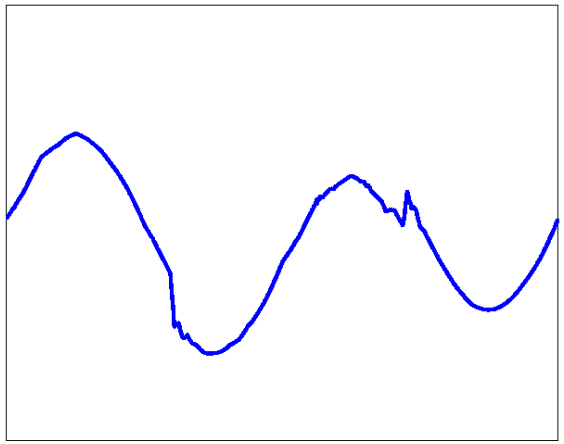
target signal



CoSaMP,
(MSE = 1.12)



L1-minimization
(MSE = 0.751)



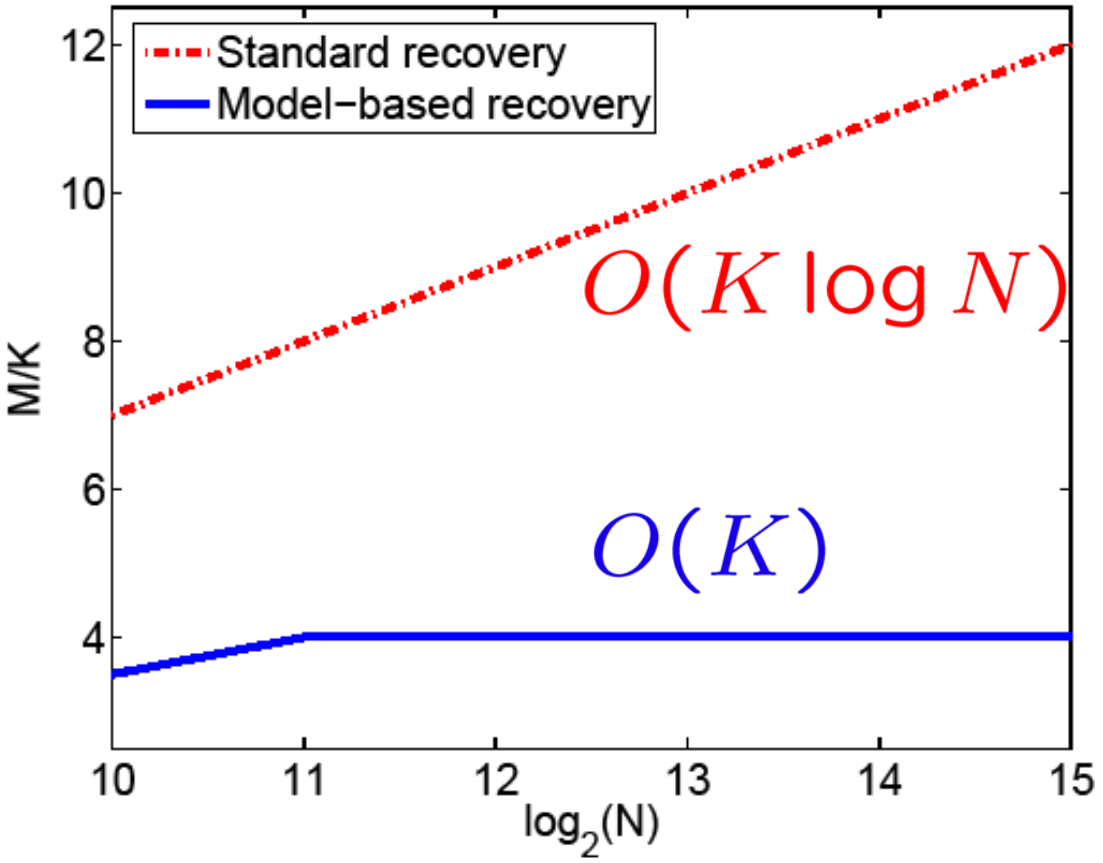
Tree-sparse CoSaMP
(MSE = 0.037)

Tree-Sparse Signal Recovery

- Number samples for correct recovery with noise

- Piecewise cubic signals + wavelets

- Plot the number of samples to reach the noise level



Clustered Sparsity

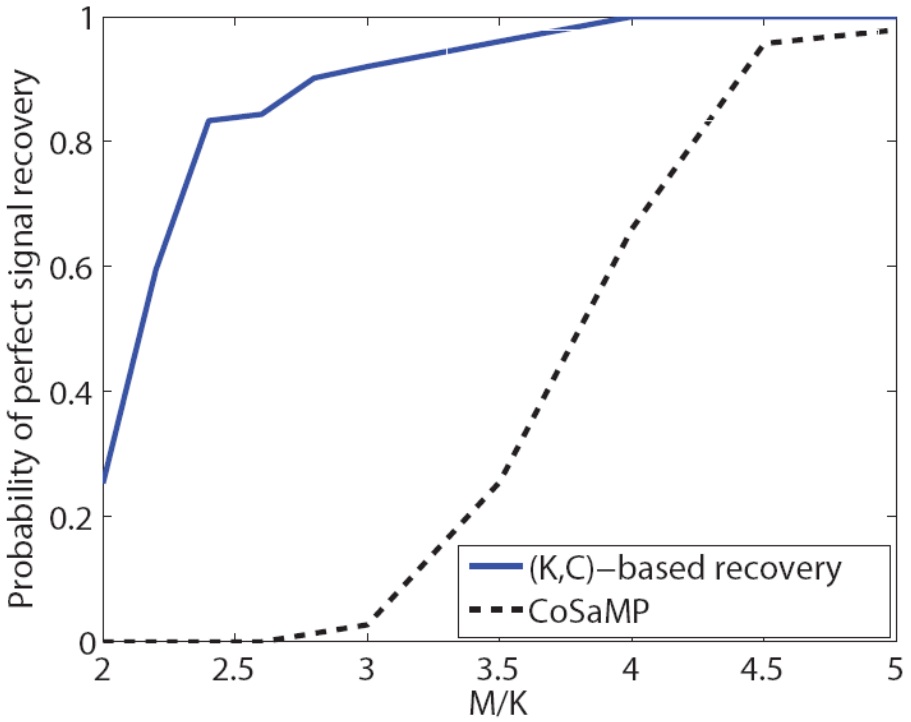
- **(*K*,*C*) sparse signals** (1-D)
  - *K*-sparse within at most *C* clusters

- For stable recovery $M = \mathcal{O}\left(K + C\log(N/C)\right)$

- Model approximation using **dynamic programming**

- Includes **block sparsity** as a special case

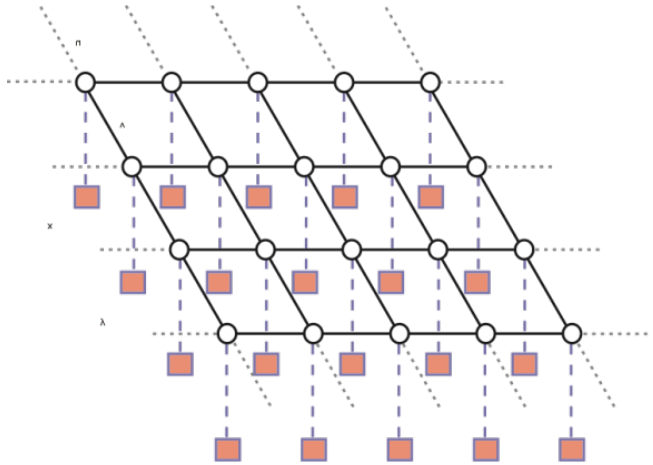
[**VC**, Indyk, Hedge, Baraniuk]

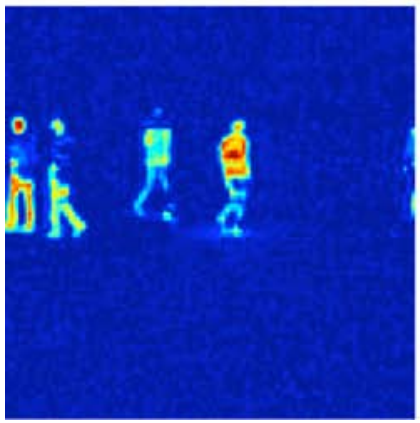


Clustered Sparsity

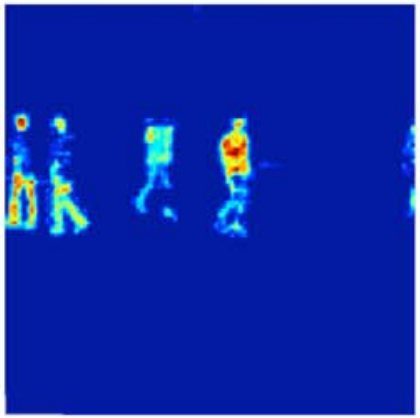
- Model clustering of significant pixels in space domain using **graphical model** (MRF)
- Ising model approximation via **graph cuts**

[**VC**, Duarte, Hedge, Baraniuk]

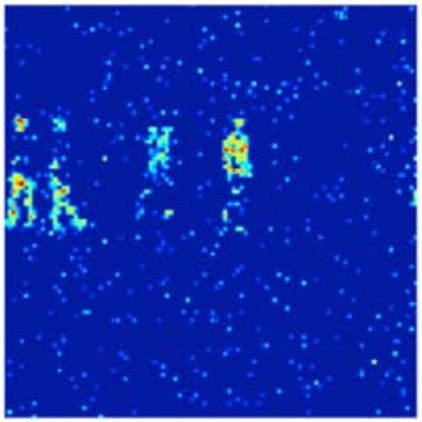




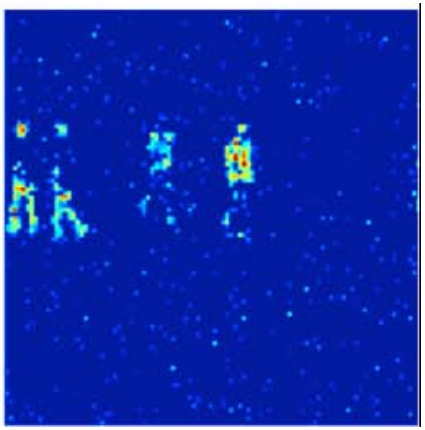
target



Ising-model recovery



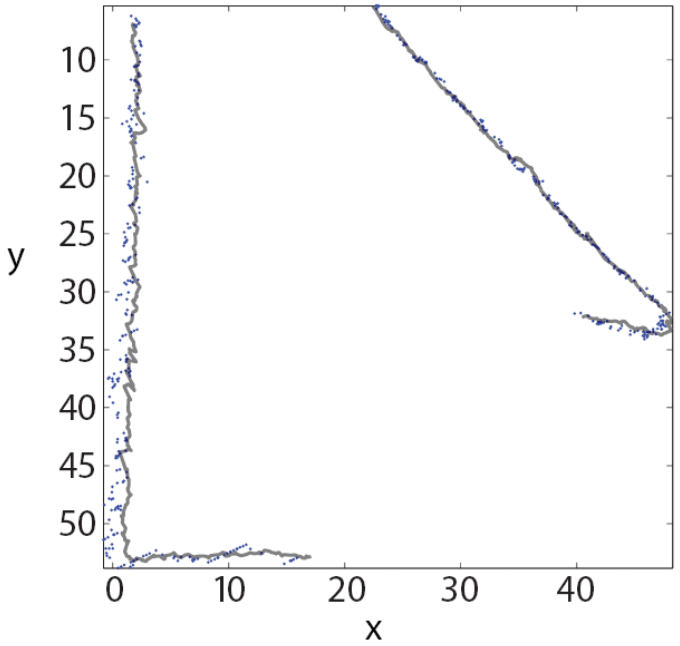
CoSaMP recovery



LP (FPC) recovery

Clustered Sparsity

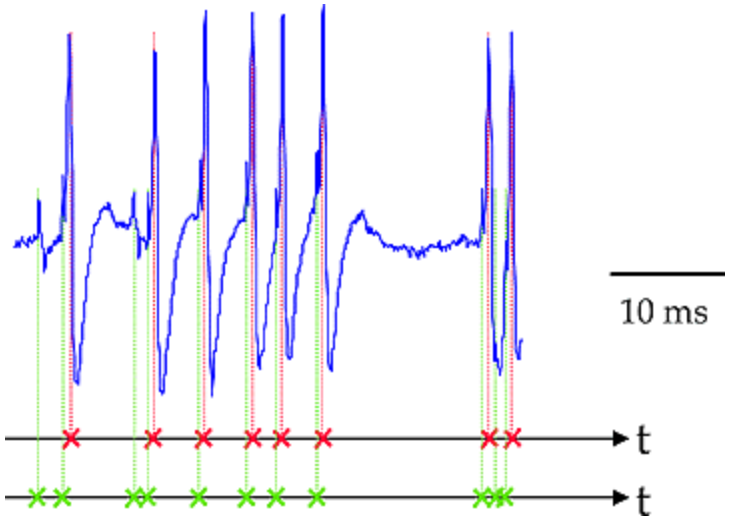
20%
Compression
No
performance
loss in tracking



Neuronal Spike Trains

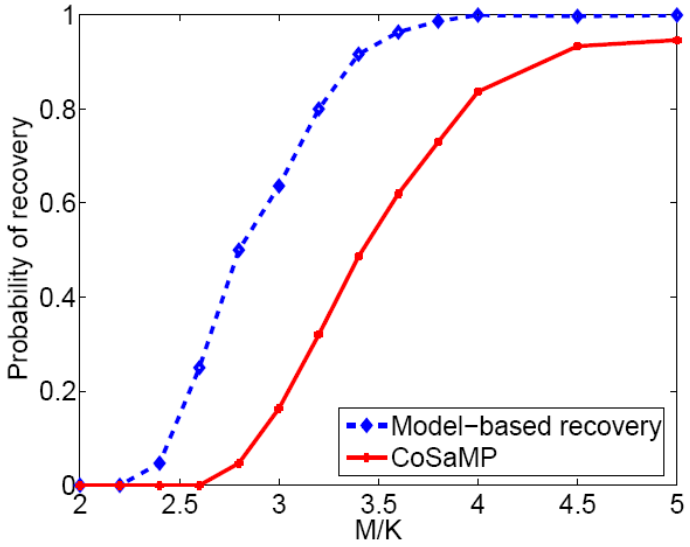
- Model the firing process of a single neuron via 1D Poisson process with spike trains
  - stable recovery

$$M = \mathcal{O}\left(K \log(N/K - \Delta)\right)$$



- Model approximation solution:
  - integer program

    efficient & provable solution due to total unimodularity of linear constraint
  - dynamic program



[Hegde, Duarte, **VC**] – best paper award

Performance of Recovery

- Using model-based IHT and CoSaMP

$$M = \mathcal{O}\left(\log |\mathcal{M}_K|\right) \qquad |\mathcal{M}_K| : \#\text{ of subspaces}$$

- **Model-sparse signals** [Baraniuk, **VC**, Duarte, Hegde]

$$\|x - \widehat{x}\|_{\ell_2} \leq C_1 \frac{\|x - x_{\mathcal{M}_K}\|_{\ell_1}}{K^{1/2}} + C_2 \|n\|_2$$

CS recovery error | signal K-term model approx error | noise

- **Model-compressible signals** w/restricted amplification property

$$\|x - \widehat{x}\|_{\ell_2} \leq C_1 \log\left(\frac{N}{K}\right) \frac{\|x - x_{\mathcal{M}_K}\|_{\ell_1}}{K^{1/2}} + C_2 \|n\|_2$$

CS recovery error | signal K-term model approx error | noise

The Probabilistic View

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Probabilistic View

- **Goal:** given $y = \Phi x + n$

  recover x

- **Prior:** iid generalized Gaussian distribution (GGD)

  iid: independent and identically distributed

$$f(x) = \mathrm{GGD}(x; q, \lambda) \propto \mathrm{e}^{-(|x|/\lambda)^q}$$

- **Algorithms:** via Bayesian inference arguments

  - maximize prior $\quad \widehat{x} = \arg\min \|x\|_q^q \text{ s.t. } y = \Phi x$
  - prior thresholding $\quad \widehat{x} = \arg\min \|y - \Phi x\|_2 \text{ s.t. } \|x\|_q \leq t$
  - maximum a posteriori (MAP) $\quad \widehat{x} = \arg\min \|y - \Phi x\|_2^2 + \mu \|x\|_q^q$

    $(\text{MAP: } n \sim \mathcal{N}(0, \sigma^2) \Rightarrow \mu = 2\sigma^2/\lambda^q)$

Probabilistic View

- **Goal:** given $y = \Phi x + n$

  recover x

- **Prior:** iid generalized Gaussian distribution (GGD)

$$f(x) = \mathrm{GGD}(x; q, \lambda) \propto \mathrm{e}^{-(|x|/\lambda)^q}$$

- **Algorithms: (q=1 <> deterministic view)**

$$\mathrm{M} = \mathrm{O}(\mathrm{K}\log(\mathrm{N}/\mathrm{K}))$$

  - maximize prior $\quad \widehat{x} = \arg\min \|x\|_q^q \text{ s.t. } y = \Phi x$
  - prior thresholding $\quad \widehat{x} = \arg\min \|y - \Phi x\|_2 \text{ s.t. } \|x\|_q \leq t$
  - maximum a posteriori (MAP) $\quad \widehat{x} = \arg\min \|y - \Phi x\|_2^2 + \mu \|x\|_q^q$

$$(\text{MAP: } n \sim \mathcal{N}(0, \sigma^2) \Rightarrow \mu = 2\sigma^2/\lambda^q)$$

Probabilistic View

- **Goal:** given $y = \Phi x + n$

  recover x

- **Prior:** iid generalized Gaussian distribution (GGD)

$$f(x) = \mathrm{GGD}(x; q, \lambda) \propto \mathrm{e}^{-(|x|/\lambda)^q}$$

- **Stable embedding:** **an experiment** by Mike Davies
  - q=1
  - x from N iid samples from GGD (no noise)
  - recover using ℓ_1

Probabilistic View

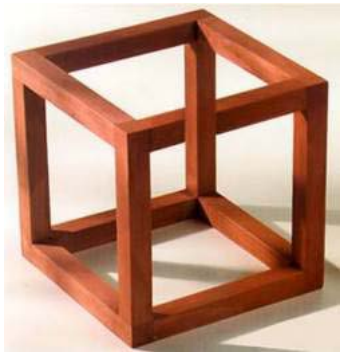
- **Goal:** given $y = \Phi x + n$

  recover x

- **Prior:** iid generalized Gaussian distribution (GGD)

$$f(x) = \mathrm{GGD}(x; q, \lambda) \propto \mathrm{e}^{-(|x|/\lambda)^q}$$

- **Stable embedding:** **a paradox**



  - q=1
  - x from N iid samples from GGD (no noise)
  - recover using ℓ_1
  - **need M~0.9 N** (Gaussian Φ)
    vs. $M = O(K \log(N/K))$

Approaches

- Do nothing / Ignore

  be content with
  where we are…

  – generalizes well

  – robust

Compressible Priors*

*You could be a Bayesian if

... your observations are less important than your prior.

[VC; Gribonval, VC, Davies; VC, Gribonval, Davies]

Compressible Priors

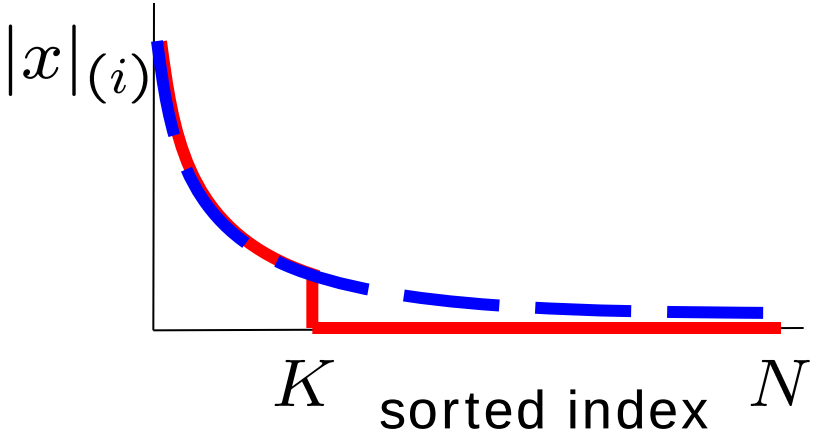
- **Goal:** seek distributions whose iid realizations $x_i \sim p(x)$ can be well-approximated as ***sparse***

Definition:

The PDF $p(x)$ is a *q-compressible prior* with parameters (ϵ, κ), when

$$\lim_{N\to\infty} \bar{\sigma}_{k_N}(x)_q \overset{a.s.}{\leq} \epsilon, \text{(a.s.: almost surely);}$$

for any sequence k_N such that $\lim_{N\to\infty} \inf \frac{k_N}{N} \geq \kappa$, where $\epsilon \ll 1$ and $\kappa \ll 1$.



relative k-term approximation:

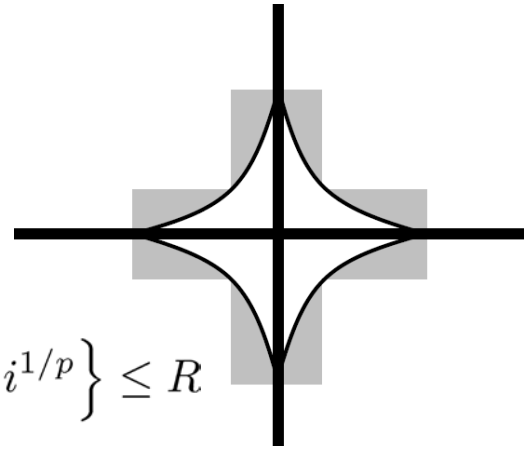
$$\bar{\sigma}_k(x)_q = \frac{\sigma_k(x)_q}{\|x\|_q}$$

$$\sigma_k(x)_q := \inf_{\|u\|_0 \leq k} \|x - u\|_q$$

Compressible Priors

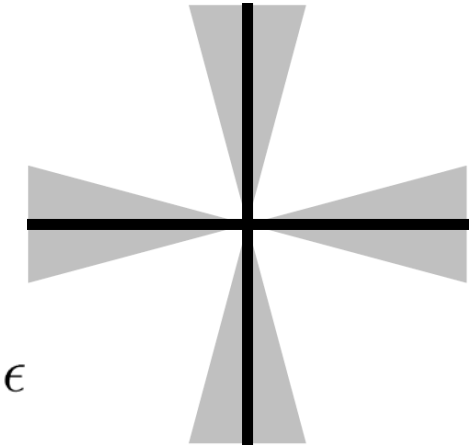
- **Goal:** seek distributions whose iid realizations can be well-approximated as ***sparse***

Classical:

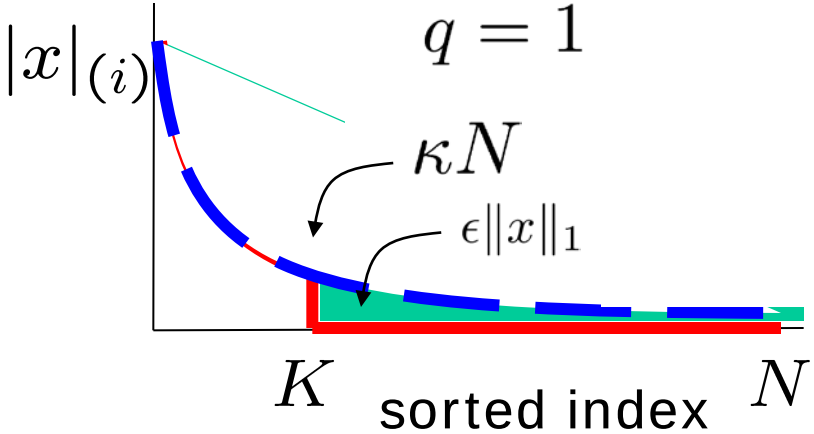


$$\|x\|_{w\ell^p} := \sup_i \left\{ |x|_{(i)} \cdot i^{1/p} \right\} \leq R$$

New:



$$\frac{\sigma_{\kappa N}(x)_q}{\|x\|_q} \leq \epsilon$$



Compressible Priors

- **Goal:** seek distributions whose iid realizations can be well-approximated as ***sparse***

- **Motivations:** deterministic embedding scaffold for the probabilistic view

  analytical proxies for sparse signals
  - learning (e.g., dim. reduced data)
  - algorithms (e.g., structured sparse)

  information theoretic (e.g., coding)

- **Main concept:** ***order statistics***

Key Proposition

Proposition 1. *Suppose* $\mathbf{x}$ *is iid with respect to* $p(x)$*. Denote* $\bar{p}(x) := 0$ *for* $x < 0$*, and* $\bar{p}(x) := p(x) + p(-x)$ *for* $x \geq 0$ *as the PDF of* $|X_n|$*, and* $\bar{F}(t) := \mathbb{P}(|X| \leq t)$ *as its cumulative distribution function. Assume that* $\bar{F}$ *is continuous and strictly increasing on some interval* $[a\ b]$*, with* $\bar{F}(a) = 0$ *and* $\bar{F}(b) = 1$*, where* $0 \leq a < b \leq \infty$*. For any* $0 \leq \kappa \leq 1$*, define the following G-function:*

$$G_q[p](\kappa) := \frac{\int_0^{\bar{F}^{-1}(1-\kappa)} x^q \bar{p}(x) dx}{\int_0^{\infty} x^q \bar{p}(x) dx}. \tag{1}$$

1. ***B****ounded moments: Let* $\mathbb{E}|X|^q < \infty$ *for some* $q \in (0, \infty)$*. Then, given any sequence* k_N *such that* $\lim_{N\to\infty} \frac{k_N}{N} = \kappa \in [0, 1]$*, the following holds almost surely*

$$\lim_{N\to\infty} \bar{\sigma}_k(\mathbf{x})_q^q \overset{a.s.}{=} G_q[p](\kappa). \tag{2}$$

2. ***U****nbounded moments: Let* $\mathbb{E}|X|^q = \infty$ *for some* $q \in (0, \infty)$*. Then, for* $0 < \kappa \leq 1$ *and any sequence* k_N *such that* $\lim_{N\to\infty} \frac{k_N}{N} = \kappa$*, the following holds almost surely*

$$\lim_{N\to\infty} \bar{\sigma}_k(\mathbf{x})_q^q \overset{a.s.}{=} G_q[p](\kappa) = 0. \tag{3}$$

Example 1

- Consider the Laplacian distribution (with scale parameter 1)

$$p_1(x) := \frac{1}{2}\exp\left(-|x|\right)$$

  - The G-function is straightforward to derive

$$G_1[p_1](\kappa) = 1 - \kappa \cdot \Big(1 + \ln 1/\kappa\Big),$$
$$G_2[p_1](\kappa) = 1 - \kappa \cdot \Big(1 + \ln 1/\kappa + \frac{1}{2}(\ln 1/\kappa)^2\Big).$$

- Laplacian distribution <> **NOT** *1 or 2-compressible*

$$\bar{\sigma}_k(\mathbf{x})_1^1 = \frac{\|x - x_K\|_1}{\|x\|_1} \leq \epsilon \Rightarrow \kappa = \frac{k_N}{N} \geq (1 - \sqrt{\epsilon})$$

Example 1

- Consider the Laplacian distribution (with scale parameter 1)

$$p_1(x) := \frac{1}{2} \exp\left(-|x|\right)$$

- Laplacian distribution <> **NOT** *1 or 2-compressible*

- **Why does ℓ_1 minimization work for sparse recovery then?**
  - The sparsity enforcing nature of the ℓ_1 cost function
  - The compressible nature of the unknown vector x

Sparse Modeling vs. Sparsity Promotion

- Bayesian interpretation of sparse recovery

four decoding algorithms:

<> **inconsistent**

$$\Delta_1(\mathbf{y}) = \underset{\tilde{\mathbf{x}}:\mathbf{y}=\mathbf{\Phi}\tilde{\mathbf{x}}}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_1,$$

$$\Delta_{\mathrm{LS}}(\mathbf{y}) = \underset{\tilde{\mathbf{x}}:\mathbf{y}=\mathbf{\Phi}\tilde{\mathbf{x}}}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_2 = \mathbf{\Phi}^{+}\mathbf{y},$$

$$\Delta_{\mathrm{oracle}}(\mathbf{y},\Lambda) = \underset{\tilde{\mathbf{x}}:\mathbf{y}=\mathbf{\Phi}\tilde{\mathbf{x}},\ \mathrm{support}(\mathbf{x})=\Lambda}{\operatorname{argmin}} \|\tilde{\mathbf{x}}\|_2 = \mathbf{\Phi}_{\Lambda}^{+}\mathbf{y},$$

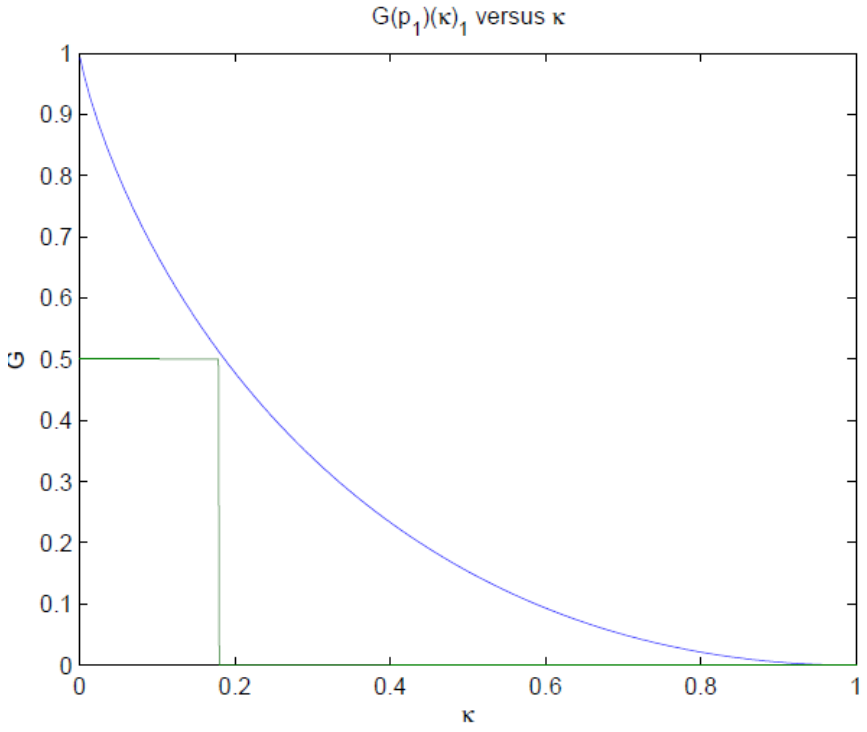
$$\Delta_{\mathrm{trivial}}(\mathbf{y}) = 0,$$

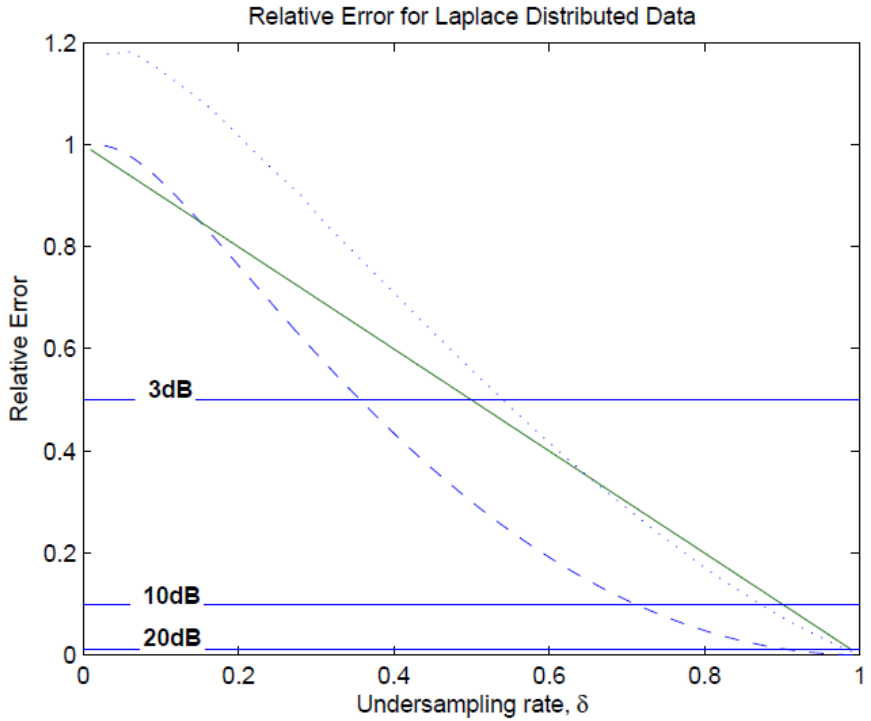
Lemma 1. *Suppose that* $\mathbf{x}$ *is iid with respect to* $p(x)$ *and that* $p(x)$ *satisfies* $G_1[p](\kappa_0) \geq 1/2$, *where* $\kappa_0 \approx 0.18$ *is an absolute constant that depends on the sensing matrix. Then, there is no undersampling ratio* $\delta = m/N$ *for which instance optimality for* Δ_1 *guarantees to outperform the trivial decoder* $\Delta_{trivial}$.

Theorem 1. *Suppose that* $\mathbf{x}$ *is iid with respect to* $p(x)$ *and that* $p(x)$ *has a finite fourth-moment* $\mathbb{E}X^4 < \infty$. *Then there exists a minimum undersampling factor* $\delta_0 = m_0/N$ *such that for any* $\delta < \delta_0$ *and any* k, *the asymptotic performance of oracle* k*-sparse estimation is almost surely worse than that of LS estimation, when* $\mathbf{n} = 0$.

($\delta_0 \approx 0.151$: Laplace)

Example 2


G(p1)(κ)1 versus κ
G
κ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8


Relative Error for Laplace Distributed Data
Relative Error
Undersampling rate, δ
3dB
10dB
20dB
0
0.2
0.4
0.6
0.8
1
1.2

Approximation Heuristic for Compressible Priors via Order Statistics

- Probabilistic signal model

$$x_i \overset{\text{iid}}{\sim} f(x) \quad (\bar{x}_i = |x_i|) \quad <> \quad \bar{x}_{(1)} \geq \bar{x}_{(2)} \geq \ldots \geq \bar{x}_{(N)}$$

order statistics of

$$\bar{f}(\bar{x}) = f(\bar{x}) + f(-\bar{x})$$

Approximation Heuristic for Compressible Priors via Order Statistics

- Probabilistic signal model

$$x_i \overset{\text{iid}}{\sim} f(x) \quad (\bar{x}_i = |x_i|) \quad <> \quad \bar{x}_{(1)} \geq \bar{x}_{(2)} \geq \ldots \geq \bar{x}_{(N)}$$

order statistics of

$$\bar{f}(\bar{x}) = f(\bar{x}) + f(-\bar{x})$$

- Deterministic signal model

$$x \in w\ell_p(R) \quad <> \quad \bar{x}_{(i)} \leq R \cdot i^{-1/p}$$

Approximation Heuristic for Compressible Priors via Order Statistics

- Probabilistic signal model

$$x_i \overset{\text{iid}}{\sim} f(x) \quad (\bar{x}_i = |x_i|) \quad <> \quad \bar{x}_{(1)} \geq \bar{x}_{(2)} \geq \ldots \geq \bar{x}_{(N)}$$

order statistics of

$$\bar{f}(\bar{x}) = f(\bar{x}) + f(-\bar{x})$$

- Deterministic signal model

$$x \in w\ell_p(R) \quad <> \quad \bar{x}_{(i)} \leq R \cdot i^{-1/p}$$

- Quantile approximation

$$\bar{x}_{(i)} \sim \mathcal{N}\left(E[\bar{x}_{(i)}], \frac{\frac{i}{N}\left(1-\frac{i}{N}\right)}{N\left[f\left(E[\bar{x}_{(i)}]\right)\right]^2} \right)$$

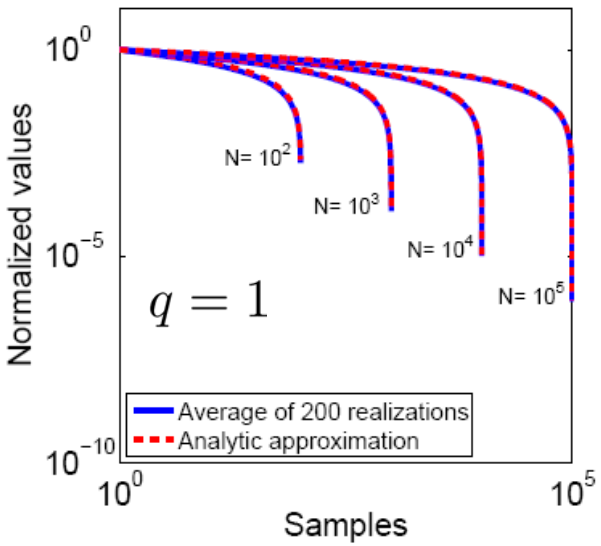
$$R = \bar{F}^{-1}\left(1 - \frac{1}{N}\right), \qquad p = R\bar{p}(R)N.$$

$$E[\bar{x}_{(i)}] = \bar{F}^{\star}\left(1 - \frac{i}{N+1}\right) \qquad \bar{F}^{\star}(u) = \bar{F}^{-1}(u)$$

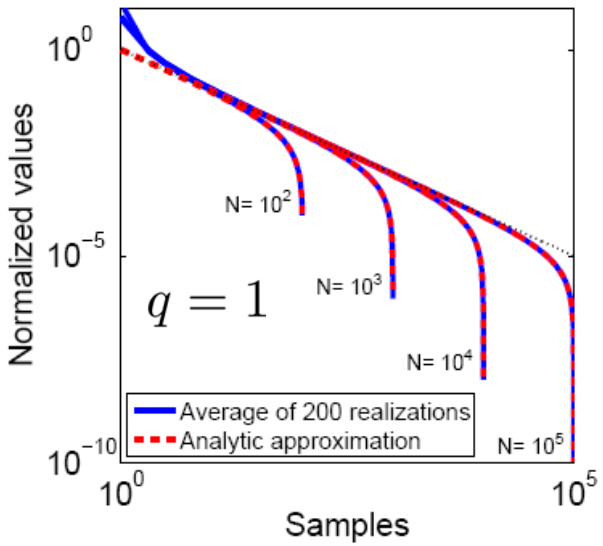
cdf

Magnitude quantile function (MQF)

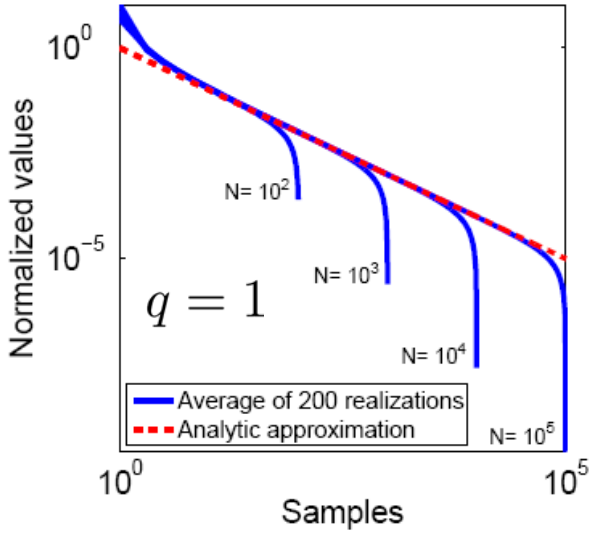
Compressible Priors w/ Examples



(a) GGD



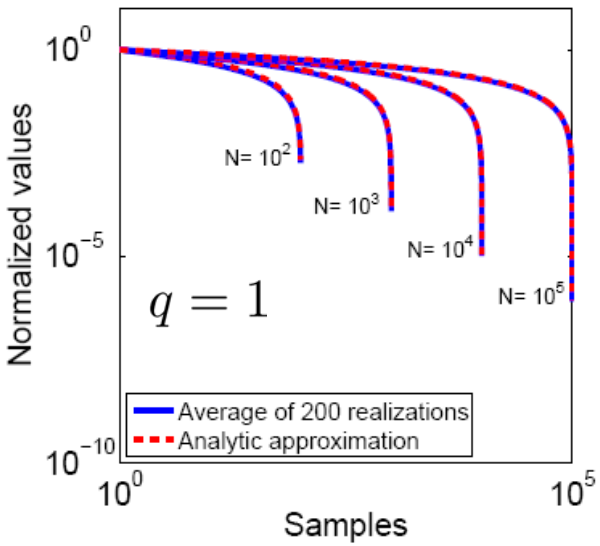
(b) GPD



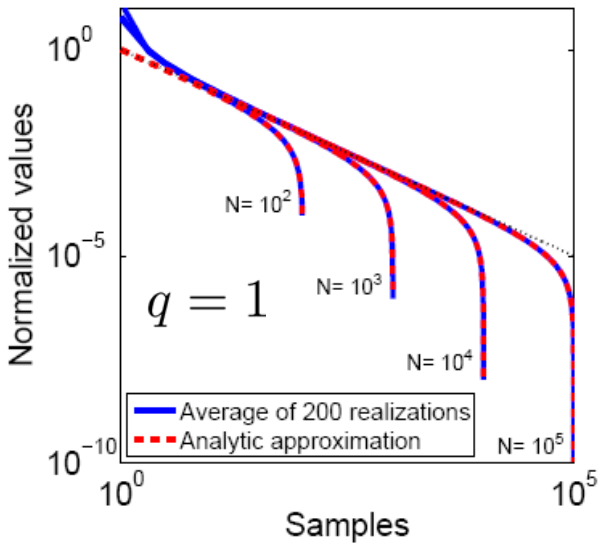
(c) Student's t

Distribution	pdf	R	p
Generalized Pareto	$\frac{q}{2\lambda}\left(1+\frac{\lvert x\rvert}{\lambda}\right)^{-(q+1)}$	$\lambda N^{1/q}$	q
Student's t	$\frac{\Gamma((q+1)/2)}{\sqrt{2\pi}\lambda\Gamma(q/2)}\left(1+\frac{x^2}{\lambda^2}\right)^{-(q+1)/2}$	$\left[\frac{2\Gamma((q+1)/2)}{\sqrt{\pi}q\Gamma(q/2)}\right]^{1/q}\lambda N^{1/q}$	q
Fréchet	$(q/\lambda)\,(x/\lambda)^{-(q+1)}\,\mathrm{e}^{-(x/\lambda)^{-q}}$	$\lambda N^{1/q}$	q
Log-Logistic	$\frac{(q/\lambda)(x/\lambda)^{q-1}}{[1+(x/\lambda)^q]^2}$	$\lambda N^{1/q}$	q
Generalized Gaussian	$\frac{q}{2\Gamma(1/q)}\mathrm{e}^{-(\lvert x\rvert/\lambda)^q}$	$\lambda\max\{1,\Gamma(1+1/q)\}\log^{1/q}(N/q)$	$q\log(N/q)$
Weibull	$(q/\lambda)\,(x/\lambda)^{q-1}\,\mathrm{e}^{-(x/\lambda)^q}$	$\lambda\log^{1/q}N$	$q\log N$
Gamma	$\frac{1}{\lambda\Gamma(q)}\,(x/\lambda)^{q-1}\,\mathrm{e}^{-x/\lambda}$	$\lambda\max\{1,\Gamma(1+1/q)^q\}\log(qN)$	$q\log(qN)$
Log-Normal	$\frac{q}{\sqrt{2\pi}x}\mathrm{e}^{-(q\log(x/\lambda))^2/2}$	$\lambda\mathrm{e}^{\sqrt{2\log N}/q}$	$\sqrt{2\log N}q$

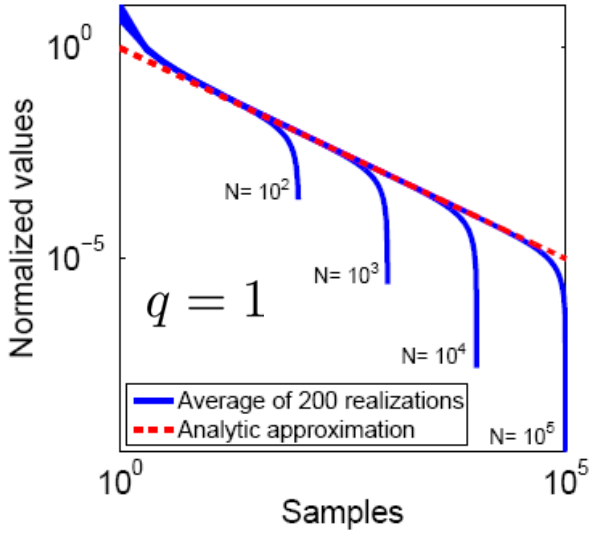
Compressible Priors w/ Examples



(a) GGD



(b) GPD



(c) Student's t

Distribution	pdf	R	p
Generalized Pareto	$\frac{q}{2\lambda}\left(1+\frac{\lvert x\rvert}{\lambda}\right)^{-(q+1)}$	$\lambda N^{1/q}$	q
Student's t	$\frac{\Gamma((q+1)/2)}{\sqrt{2\pi\lambda}\Gamma(q/2)}\left(1+\frac{x^2}{\lambda^2}\right)^{-(q+1)/2}$	$\left[\frac{2\Gamma((q+1)/2)}{\sqrt{\pi q}\Gamma(q/2)}\right]^{1/q}\lambda N^{1/q}$	q
Fréchet	$(q/\lambda)\,(x/\lambda)^{-(q+1)}\,\mathrm{e}^{-(x/\lambda)^{-q}}$	$\lambda N^{1/q}$	q
Log-Logistic	$\frac{(q/\lambda)(x/\lambda)^{q-1}}{[1+(x/\lambda)^q]^2}$	$\lambda N^{1/q}$	q
Generalized Gaussian	$\frac{q}{2\Gamma(1/q)}\mathrm{e}^{-(\lvert x\rvert/\lambda)^q}$	$\lambda\max\{1,\Gamma(1+1/q)\}\log^{1/q}(N/q)$	$q\log(N/q)$
Weibull	$(q/\lambda)\,(x/\lambda)^{q-1}\,\mathrm{e}^{-(x/\lambda)^q}$	$\lambda\log^{1/q}N$	$q\log N$
Gamma	$\frac{1}{\lambda\Gamma(q)}\,(x/\lambda)^{q-1}\,\mathrm{e}^{-x/\lambda}$	$\lambda\max\{1,\Gamma(1+1/q)^q\}\log(qN)$	$q\log(qN)$
Log-Normal	$\frac{q}{\sqrt{2\pi}x}\mathrm{e}^{-(q\log(x/\lambda))^2/2}$	$\lambda\mathrm{e}^{\sqrt{2\log N}/q}$	$\sqrt{2\log N}q$

Dimensional (in)Dependence

- **Dimensional independence** $p = p(\theta)$ **< >** $M = O(K \log(N/K))$

< > unbounded moments

example: $K = (p/\epsilon)^{\frac{p}{1-p}} \Rightarrow \|x - x_K\|_1 \leq \epsilon \|x\|_1$

$$\|x - \widehat{x}\|_2 \leq C_1 \frac{\|x - x_K\|_1}{K^{1/2}} + C_2 \|n\|_2$$

CS recovery error

signal K-term approx error

noise

$$M = O(K \log(N/K))$$

Dimensional (in)Dependence

- **Dimensional independence** $p = p(\theta)$ **<>** $M = O(\log N)$

$$K = (p/\epsilon)^{\frac{p}{1-p}} \Rightarrow \|x - x_K\|_1 \leq \epsilon \|x\|_1$$

truly logarithmic embedding

- **Dimensional dependence** $p = p(\theta, N)$ **<>** $M = o(N)$

<> bounded moments

example: iid Laplacian OS: $\bar{x}_{(i)} \approx \lambda \log \frac{N}{i}$

$$K = (1 - \sqrt{\epsilon})N \Rightarrow \|x - x_K\|_1 \leq \epsilon \|x\|_1$$

not so much! / same result can be obtained via the G-function

Why should we care?

- **Natural images**
  - wavelet coefficients

deterministic view vs. **probabilistic view**

Besov spaces — GGD, scale mixtures

wavelet tresholding — Shannon source coding

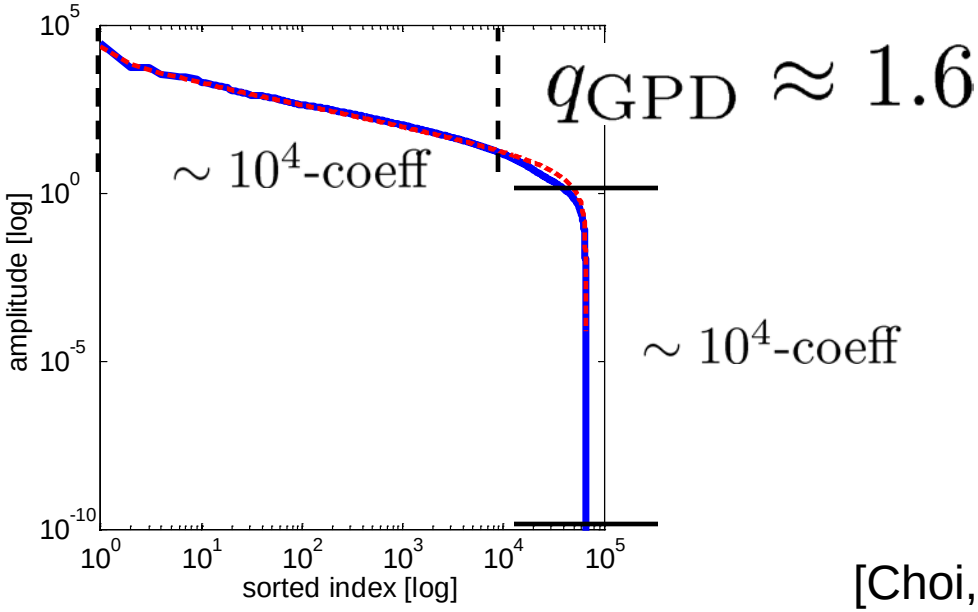
Why should we care?

- **Natural images**
  - wavelet coefficients

deterministic view vs. **probabilistic view**

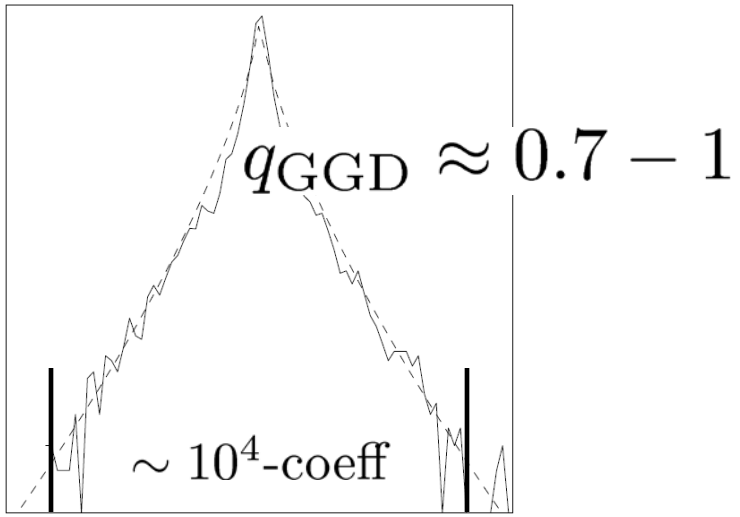
Besov spaces

wavelet tresholding



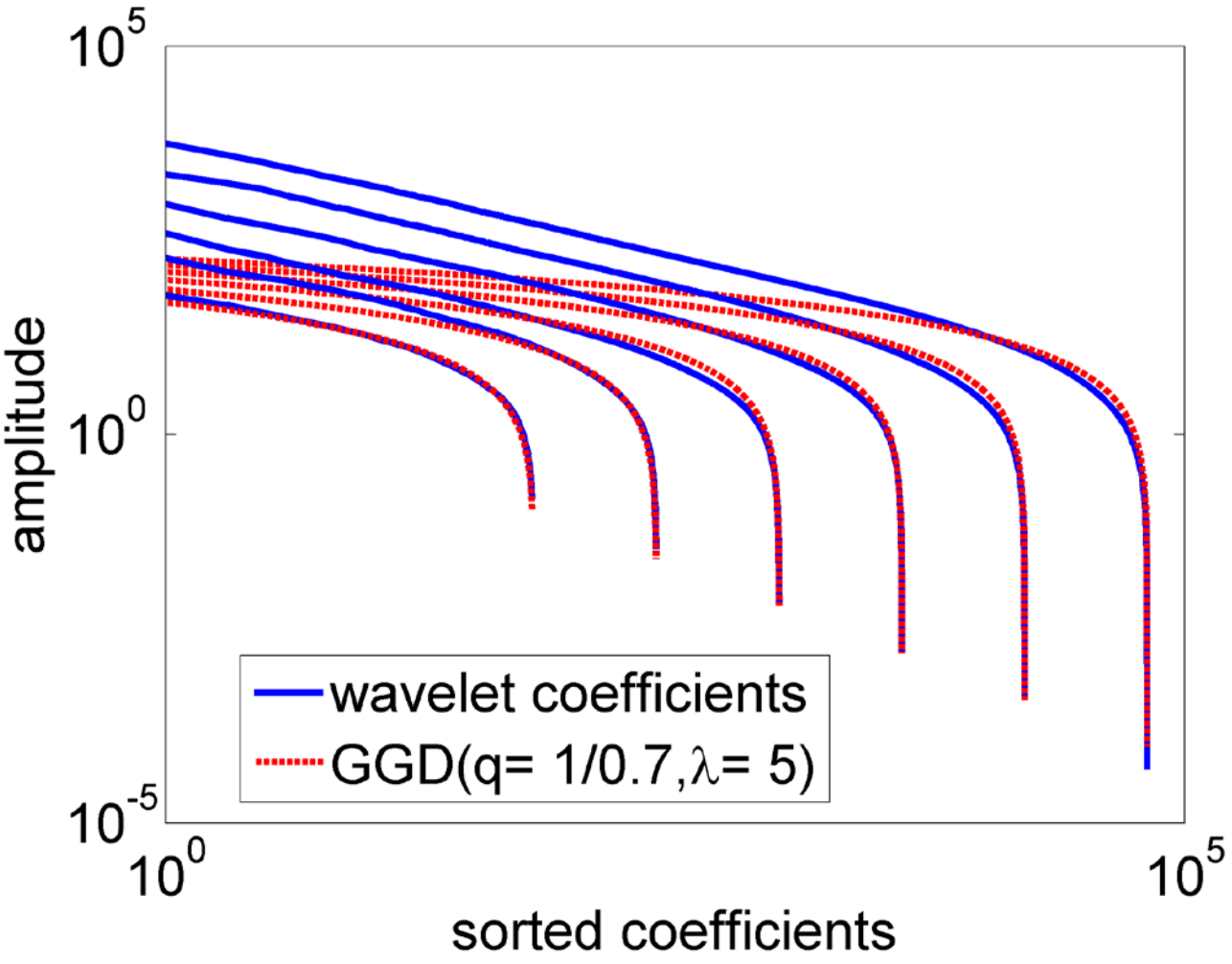
GGD, scale mixtures

Shannon source coding

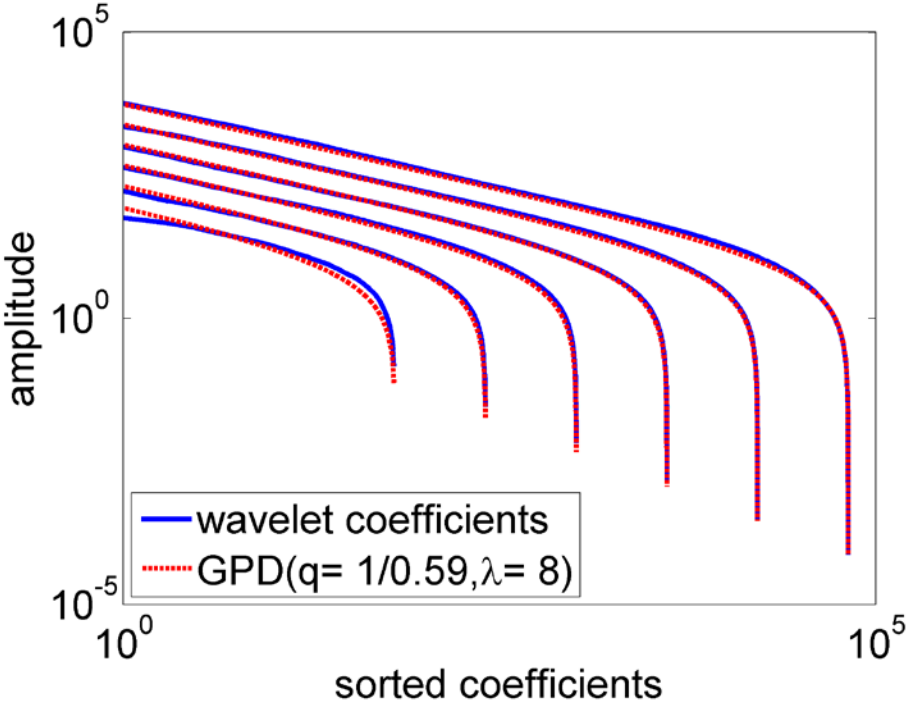


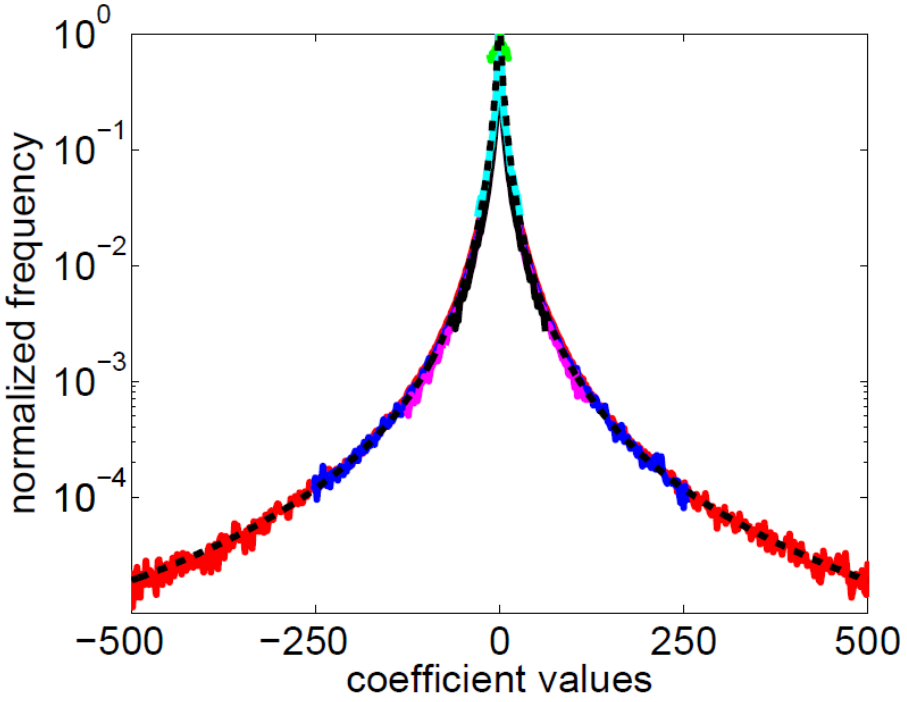
[Choi, Baraniuk; Wainwright, Simoncelli; …]

Berkeley Natural Images Database



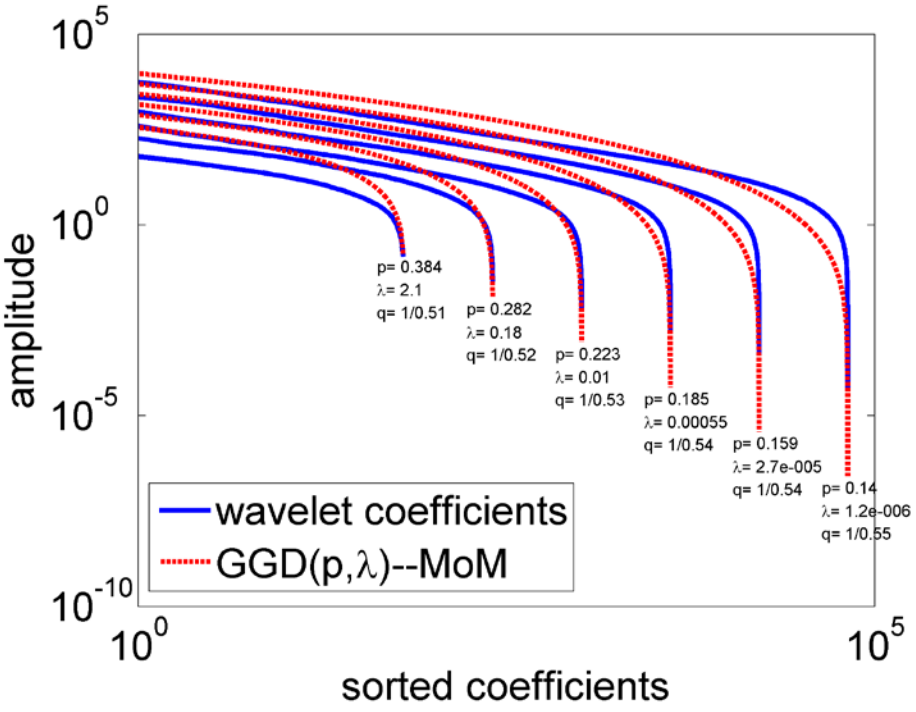
Berkeley Natural Images Database

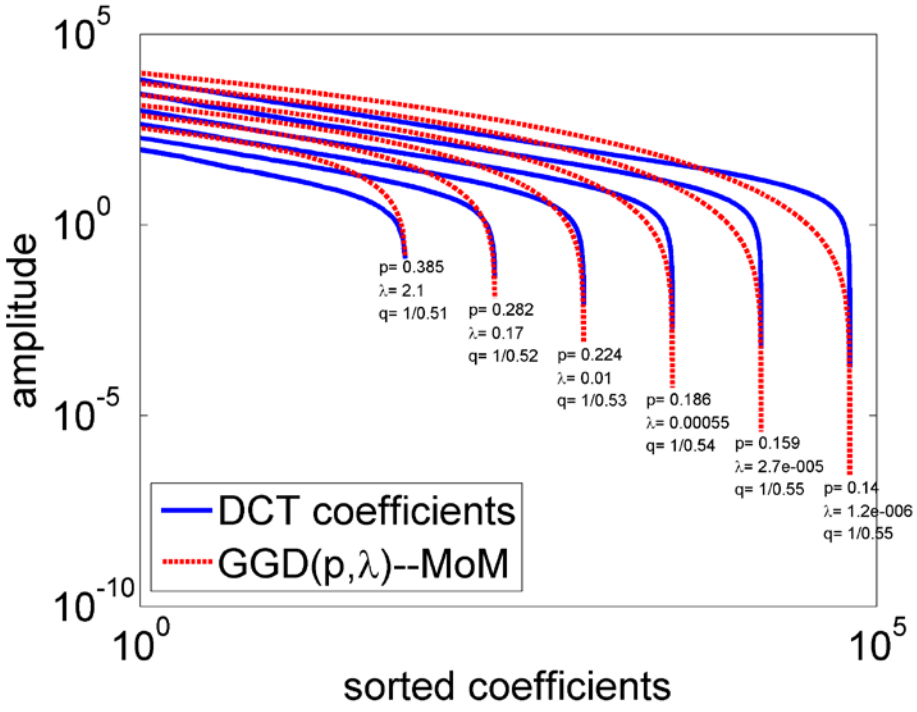




$$\log \mathrm{GPD}(q,\lambda) \doteq -(q+1)\log\left(1+\frac{|x|}{\lambda}\right) \approx -\frac{|x|}{\lambda/(q+1)}$$

Berkeley Natural Images Database





Learned parameters depend on the dimension

Why should we care?

- **Natural images** (coding / quantization)
  - wavelet coefficients

deterministic view vs. **probabilistic view**

Besov spaces

wavelet tresholding

GGD, scale mixtures

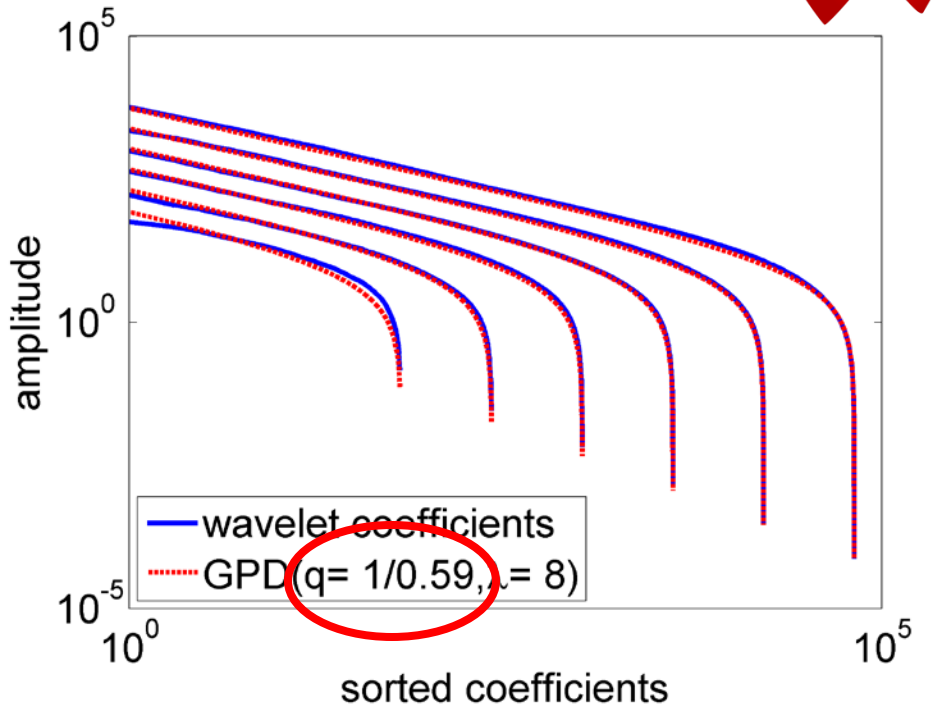
Shannon source coding

(histogram fits, KL divergence)

[bad ideas]

- **Conjecture:** Wavelet coefficients of natural images belong to a dimension independent (non-iid) compressible prior

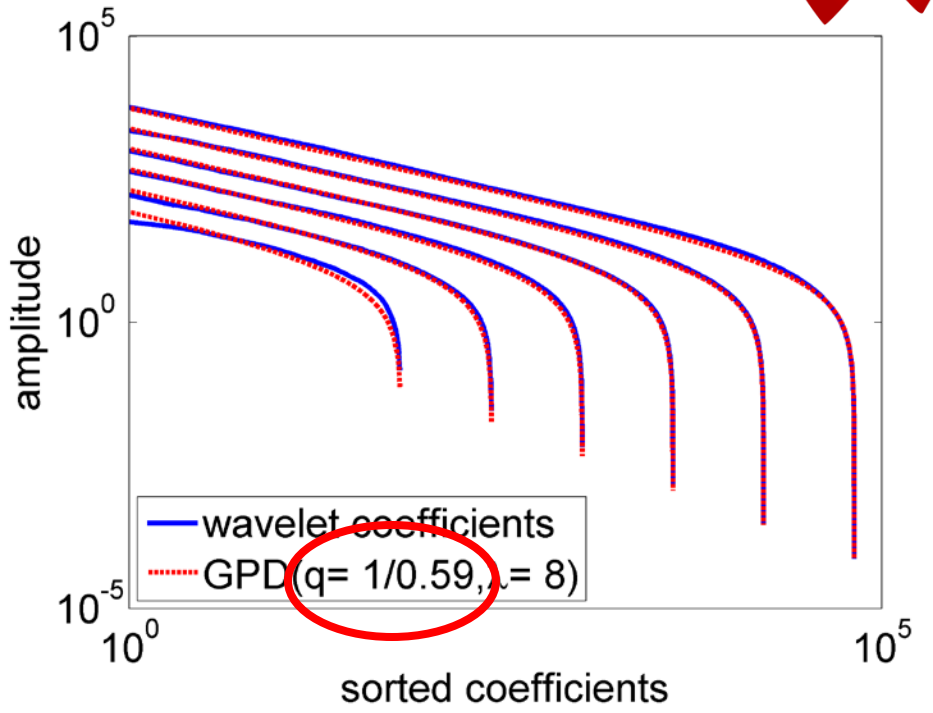
Incompressibility of Natural Images



1-norm instance optimality blows up:

$$\|x - \widehat{x}\|_2 \leq C_1 \frac{\|x - x_K\|_1}{K^{1/2}} + C_2 \|n\|_2$$

Incompressibility of Natural Images



1-norm instance optimality blows up:

$$\|x - \widehat{x}\|_2 \leq C_1 \frac{\|x - x_K\|_1}{K^{1/2}} + C_2 \|n\|_2$$

Is compressive sensing USELESS for natural images?

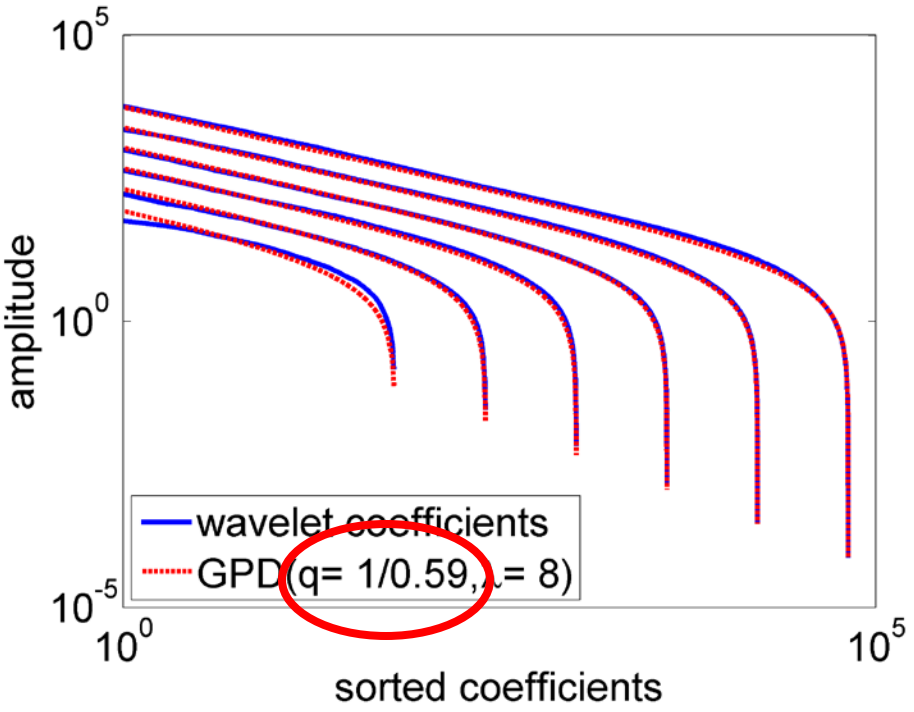
Instance Optimality in Probability to the Rescue

Theorem 2 (Asymptotic performance of the ℓ_1 decoder under infinite second moment)**.** *Let $X_n, n \in \mathbb{N}$ be iid samples from a distribution with PDF $p(x)$ satisfying the hypotheses of Proposition 1. Assume that $\mathbb{E}X^2 = \infty$, and define the coefficient vector $\mathbf{x}_N = (X_1, \ldots, X_N) \in \mathbb{R}^N$. Similarly let $\phi_{i,j}$, $i, j \in \mathbb{N}$ be iid Gaussian variables $\mathcal{N}(0,1)$ and define the $m_N \times N$ Gaussian random matrix $\mathbf{\Phi}_N = \left[\phi_{ij}/\sqrt{m_N}\right]_{1 \le i \le m_N, 1 \le j \le N}$.*

Consider a sequence of integers m_N such that $\lim_{N\to\infty} m_N/N = \delta$ then

$$\frac{\|\Delta_1(\mathbf{\Phi}_N \mathbf{x}_N) - \mathbf{x}_N\|_2}{\|\mathbf{x}_N\|_2} \xrightarrow{a.s.} 0$$

Incompressibility of Natural Images

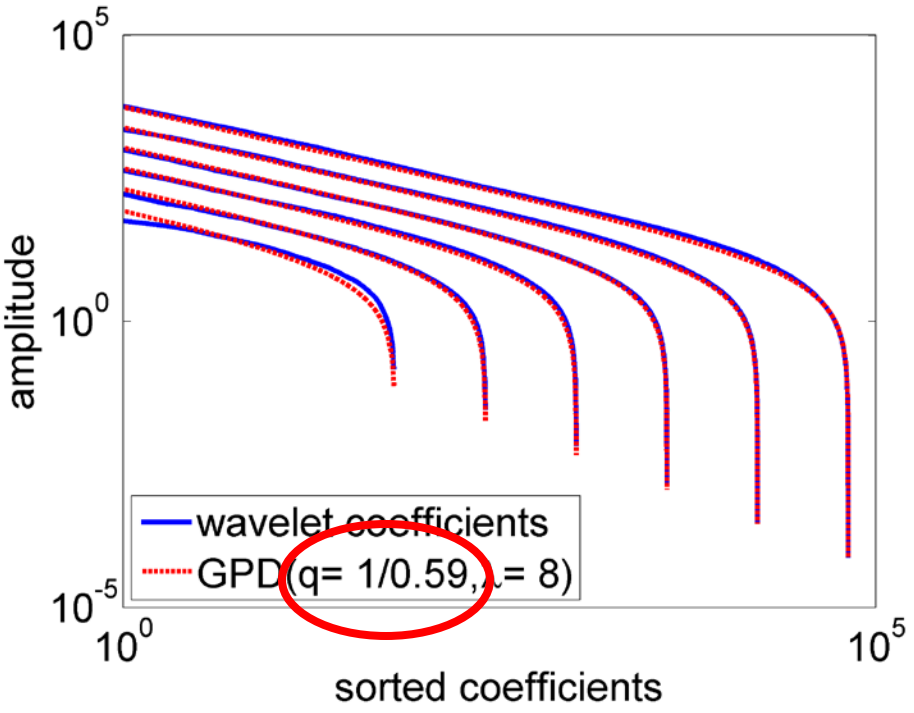


Is compressive sensing
USELESS
for natural images?

Not according to Theorem 2!!!

For large *N*, 1-norm minimization is still near-optimal.

Incompressibility of Natural Images

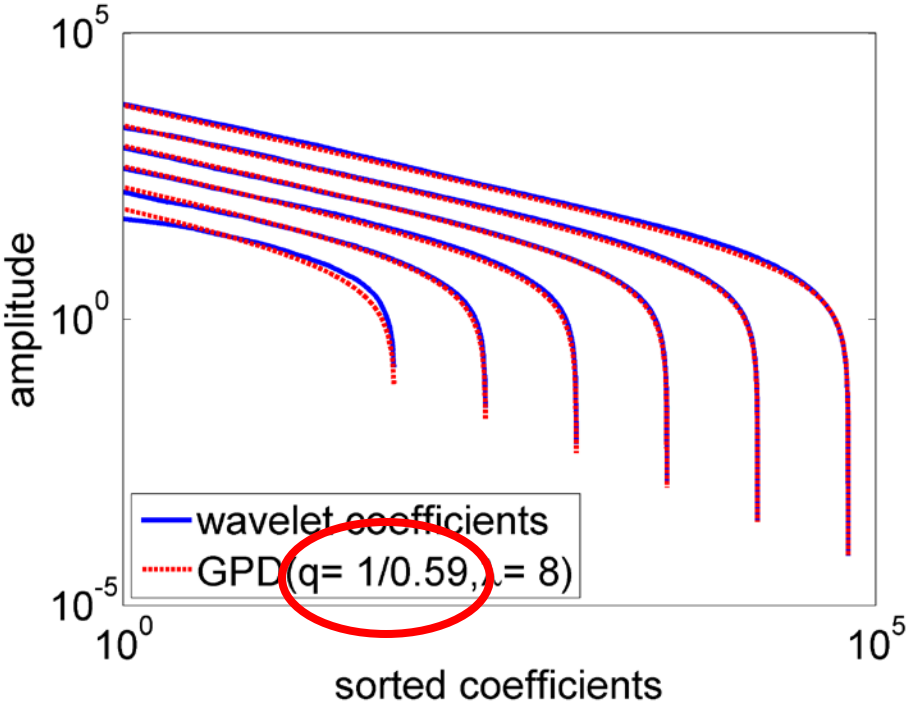


Is compressive sensing USELESS for natural images?

Not according to Theorem 2!!!

But, are we NOT missing something practical?

Incompressibility of Natural Images



But, are we NOT missing something practical?

Natural images have finite energy since we have finite dynamic range.

While the resolution of the images are currently ever-increasing, their dynamic range is not.

In this setting, compressive sensing using naïve sparsity will not be useful.

Other Bayesian Interpretations

- Multivariate Lomax dist.
  (non-iid, compressible w/ r=1)
  $\lambda_i = \lambda$

$$f(x_1, \ldots, x_N) \propto \frac{1}{\left(1+\sum_i \lambda_i^{-1}|x_i|\right)^{q+N}}$$

(has GPD(x_i;q,λ_i) marginals)

  - maximize prior

$$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } y = \Phi x$$

  - prior thresholding

$$\widehat{x} = \arg\min \|y - \Phi x\|_2 \text{ s.t. } \|x\|_1 \leq t$$

  - maximum a posteriori (MAP)

$$\widehat{x}^{\{k\}} = \arg\min \|y - \Phi x\|_2^2 + \mu^{\{k\}}\|x\|_1$$

$$(n \sim \mathcal{N}(0, \sigma^2) \Rightarrow \mu^{\{k\}} = 2\sigma^2(q+N)/(\lambda + \|\widehat{x}^{\{k-1\}}\|_1))$$

fixed point continuation

- Interactions of Gamma and GGD

$$f(x) \propto \frac{1}{(1+|x|^r/\lambda^r)^{\frac{q+1}{r}}}$$

  - iterative re-weighted ℓ_r algorithms

Summary of Results

Table 1: Simple Rule of Thumbs for IID Compressibility and Linear Regression

Moment property	$\mathbb{E}x^2 = \infty$	$\mathbb{E}x^2 < \infty$ and $\mathbb{E}x^4 = \infty$	$\mathbb{E}x^4 < \infty$
General result	Δ_1 performs ideally for any δ	N/A depends on finer properties of $p(x)$	Δ_{LS} outperforms Δ_{oracle} for small $\delta < \delta_0$
Examples		Example: $p_0(x) := 2\|x\|/(x^2+1)^3$ Δ_{oracle} performs just as Δ_{LS}	Example: $p_{\tau,\lambda}(x) \propto \exp\left(-\|x/\lambda\|^\tau\right)$ $0 < \tau < \infty$ Generalized Gaussian
	Example: $p(x) \propto (1+\|x/\lambda\|^\tau)^{-(q+1)/\tau}$ Generalized Pareto ($\tau = 1$) / Student's t ($\tau = 2$)		
	Case $0 < q \leq 2$	Case $2 < q < 4$ Δ_{oracle} outperforms Δ_{LS} for small $\delta < \delta_0$	Case $q > 4$

$\delta = M/N$