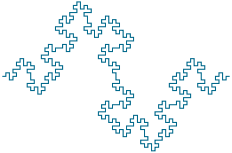
Some Aspects of Sparse Representations

Theja Tulabandhula
Massachusetts Institute of Technology



Sourced from Candes, Tao, Romberg, Baranuik, Tropp, Rosasco, Recht *and* a lot of other material.

February 10, 2011

INTRODUCTION

INTRODUCTION

SPARSITY

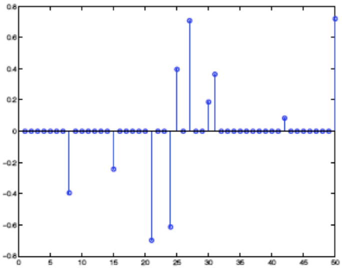
COMPRESSIVE SAMPLING

KEY POINTS

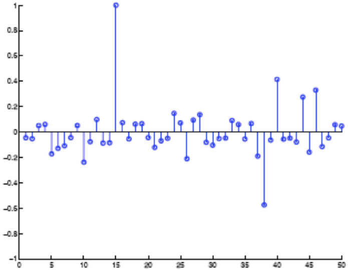
1. Sparsity is a very interesting pattern to exploit in data analysis.
2. Compressed Sensing uses sparsity.
3. It is an alternative framework for sampling : sample selectively and compress at the same time.
4. Work on decompressing by solving an optimization problem.
5. Very applicable when sensors expensive or when acquisition time of signals is an issue.
6. Different from classical Nyquist theory where one sample everything and then does compress.

THE SPARSITY HEURISTIC

1. Sparse means signal has fewer degrees of freedom than its nominal dimension.
2. Def: $x \in R^n$ is s-sparse if exactly s components of x are non-zero.
3. Def: Support of x: $\text{supp}(x) = \{i : x_i \neq 0\}$.
4. Def: Cardinality of x: $\text{card}(x) = |\text{supp}(x)|$. Also called ℓ_0 norm.
5. Researchers assume some structure of the problem at hand. An assumption.
6. Sparsity is a structure.
7. It is the idea that signals can be expressed with few coefficients of an (orthogonal) basis.



Sparse signal



Nearly sparse signal

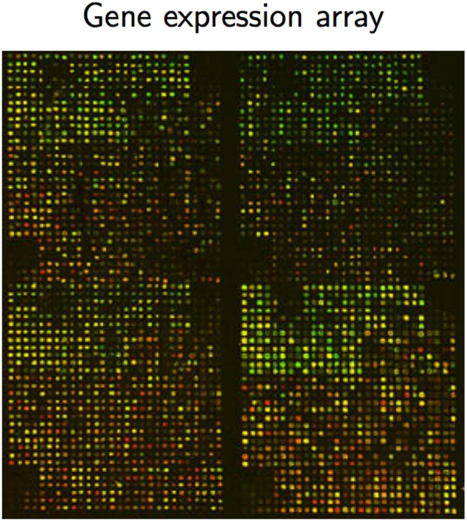


Figure: Sparsity in multi-dimensional data

SPARSEST VECTOR PROBLEM

1. Given $\epsilon \geq 0$, matrix $A_{m \times n}$ and $b \in R^m$, what is the sparsest x satisfying $||Ax - b|| \leq \epsilon$?
2. That is,

$$\begin{aligned} \min \operatorname{card}(x) \\ \text{subject to } ||Ax - b|| \leq \epsilon \end{aligned}$$

3. Special interesting case: $\epsilon = 0$.
4. $m < n$: we will have an underdetermined system here.

REASONS FOR FINDING A LOW CARDINALITY SOLUTIONS

1. **Model selection**: Consider n scalar inputs$\{x_i\}$ which affect an outcome y.
   - Example: indicator variable for “”Rajon Rondo is playing tonight” or indicator variable for “past three matches won” and so on.
   - Goal: Predict outcome (Celtics will win or lose) from these inputs.
   - Model: $y = \sum_i w_i x_i$.
   - Secondary goal: Minimize card(x).
2. **Astronomy**: A picture of night time sky: Is sparse. Only a few pixels bright.
3. **Error correction codes**: Error vectors which get ‘added’ to signal vectors are sparse.

ONE OTHER EXAMPLE

Compressive sampling: Most of the signals are sparse in some basis. Take a few coded samples from which one can recover the signal exactly.

*We will localize to this topic in a few slides

NP HARDNESS

1. Finding a sparse vector that approximates $Ax = b$ is NP hard. Reduction from Exact cover problem (Natarajan 95)
2. Exact cover problem: Set $S = \{s_1, s_2, ..., s_m\}$. $|S| = m$. C collection of n subsets of S each having 3 elements. Does there exist a sub-collection $\hat{\text{C}}$ of C such that every element of S is a member of exactly one set in $\hat{\text{C}}$?
3. Let $C = \{C_1, C_2, ..., C_n\}$. Define $m \times n$ matrix with $A_{i,j} = \begin{cases} 1 & s_i \in C_j \\ 0 & \text{otherwise} \end{cases}$
4. A has 3 ones per column. $b = [11 \ldots 1]^T, b \in R^m$.

   Proposition: If we have an algorithm for sparsest vector problem and it returns x with card(x) = s. Then card$(x) \leq \frac{m}{3}$ if and only if an exact cover exists.

SPARSITY AS A REGULARIZER

1. Regularize: Incorporate a prior in a learning task
2. Given input output pairs $\{x, y\}_1^m$ sampled i.i.d from a distribution, want to find a good predictor (e.g, regressor) for future inputs
3. Obtain by minimizing

$$\min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \text{loss}(f(x^i), y^i) + \gamma ||f||$$

4. $||f||$ is the regularization term. $\gamma > 0$ determines strength of regularization.
5. For linear input output models, f is parametrized by vector w.
6. Choose $||f||$ = card(w).

Sparsity as regularizer

1. Benefits: Interpretability of the model, compression.
2. Enforcing sparsity may avoid overfitting. But if we select too few inputs as relevant, then might underfit.
3. This regularized problem is again NP-hard.
4. So we approximate: Based on convex relaxation and greedy methods.

   Keywords: **Basis Pursuit** and **Lasso**.

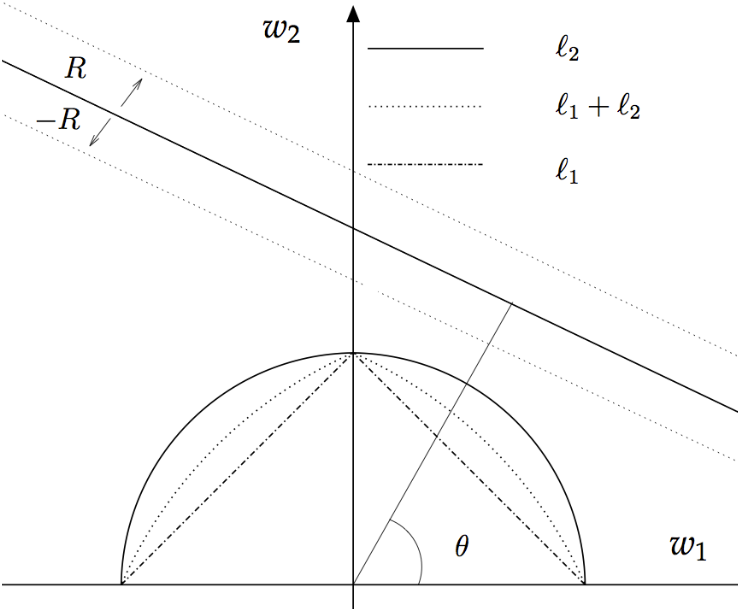
CONVEX SUBSTITUTE FOR CARD(x)

1. Use an ℓ_1 norm instead of ℓ_0 or card$(\cdot)$.
2. Is there an intuition why this substitution works and ℓ_2 norm (or Tikhonov regularizer) substitution doesn't?
3. Reformulate optimization problem as

$$\min_w \sum_{i=1}^{n} |w_i|$$

$$\text{subject to } \frac{1}{m}\sum_{i=1}^{m} \text{loss}(f(x^i), y^i) \leq R$$

4. Choose loss to be squared loss. Let $Y = [y^1, \ldots, y^m]^T$ and $X = [x^1 \ldots x^m]^T$. Loss can be rewritten as $\frac{1}{m}||Y - Xw||_2^2$.


w_2
ℓ_2
$\ell_1 + \ell_2$
ℓ_1
R
$-R$
θ
w_1

1. Solve the learning task using simple iterative algorithms.
2. Simple iterative thresholding algorithm.
3. Let $w^0 = 0$. At iteration k, $w^k = S[w^{k-1} + \tau X^T(Y - Xw^{k-1})]$.
4. where τ iteration constant, $S[]$ is a soft thresholding function.
5. Non-unique solutions in general.

SPARSITY AND ENTROPY

1. Consider binary sequences in $\{0-1\}^n$ with exactly s ones.
2. Need at least $\binom{\log_2(n}{s)}$ bits (entropy).
3. Stirling's approximation this is $\sim s \log \frac{n}{s}$.
4. Thus when $s << n$, entropy is small.

SPARSITY FOR COMPRESSION (SOURCE CODING!)

1. Let

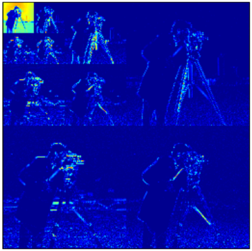
$$f = \sum_{j=1}^{n} x_j \psi_j$$

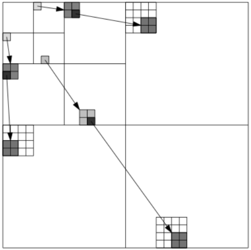
$$\text{or } f = \Psi x$$

2. x is the coefficient sequence. Ψ can be diracs, sinusoids (fourier) or even just the canonical basis in R^n.
3. A scheme for compression:
   - Choose s.
   - Keep the s terms with largest coefficients $x_s = \sum_{j \in J} \psi_j x_j : \#(J) = s$.
4. Few coefficients can capture a lot of energy of the signal. Discard the low energy components (lossy compression technique)
5. Most transform coders (wavelets!) make use of this feature while compressing images.

Wavelets for compression

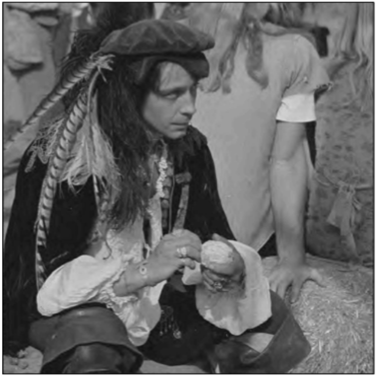
1. Near-sparse representations are well approximated by sparse representations
2. This ideas is used in transform coding: DCT, Wavelets (image compression).





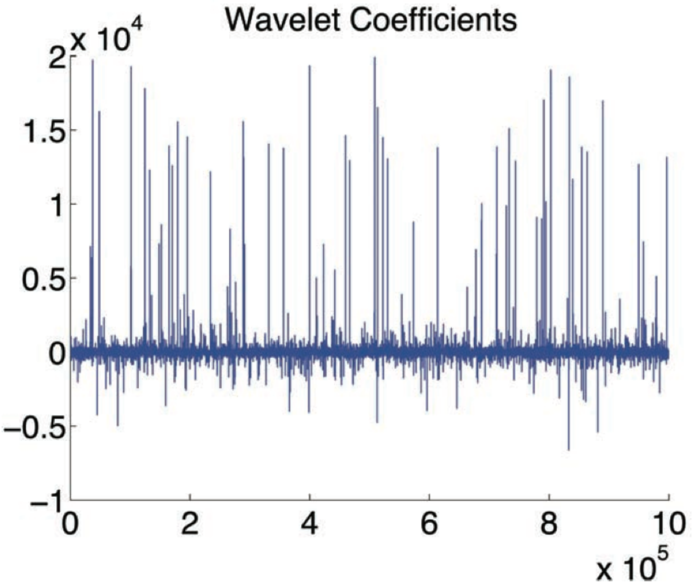
WAVELETS FOR COMPRESSION

1 megapixel image



25k term approx

WAVELETS FOR COMPRESSION



HISTORY: SPARSITY IN PRACTICE

- Geology/geophysics
  - Claerbout and Muir (1973)
  - Taylor et al. (1979)
  - Levy and Fullager (1981)
  - Oldenburg et al. (1983)
  - Santosa and Symes (1988)
- Radio astronomy
  - Högbom (1974)
  - Schwarz (1978)
- Fourier transform spectroscopy
  - Kawata et al. (1983)
  - Mammone (1983)
  - Minami et al. (1985)
- NMR spectroscopy
  - Barkhuijsen (1985)
  - Newman (1988)
- Medical ultrasound
  - Papoulis and Chamzas (1979)

Common elements:

- Sparsity assumption enables improved resolution of estimate (beyond bandwidth of acquisition)
- Sparsity in space or gradient with respect to space
- ℓ_1 minimization to promote sparsity
- Sparse domain given by nature

Mathematical limits explored in [Donoho (1992)]

HISTORY: SPARSITY IN PRACTICE

1. Matching Pursuit 1993
2. LASSO (Tibshirani) 1996
3. Basis Pursuit 1998

   .

   .

   .

4. Compressive sensing (also known as Compressed sensing, Compressive sampling, Sketching and Sparse sampling): Donoho 2004, Candes-Tao 2004.

CLASSICAL NYQUIST THEORY

1. Conventional wisdom in signal acquisition and reconstruction from frequency: the number of fourier samples must match the desired accuracy or resolution (for example, number of pixels)
2. f is bandlimited if $\forall |\omega| > \Omega/2\pi, \hat{f}(\omega) = 0$. Bandwidth = Ω/π
3. Classical Nyquist: f can be reconstructed from its time samples, sampled at frequency $\geq \Omega/\pi$,

$$f(t) = \sum_{n \in \mathbb{Z}} f(n\pi/\Omega) sinc(\Omega t - n\pi)$$

4. Exact reconstruction with samples spaced π/Ω apart i.e., sampled at frequency Ω/π samples/second.
5. If Ω/π is too large, the (analog to digital) sampling hardware will need to work really fast!

PROBLEM AT HAND

1. Wish to obtain $f \in \mathbb{R}^n$ from m measurements.
2. Can look at it as asking m non-adaptive queries (inner-products) about f.

$$y_k = \varphi_k^T f, k = 1, ..., m.$$

$$\text{equivalently } Y_{m\times 1} = \Phi_{m\times n} f_{n\times 1}.$$

3. Example, for an tradition digital camera, φ_k is just an indicator function.
4. Example, for traditional Nyquist sampling of a scalar time domain signal, $\varphi_k, k = 1, ..., m(= n)$ is just a unit vector (canonical basis in R^n).
5. Is it possible to take only $m << n$ measurements and get all information about f?

Motivation for the problem

1. There is cost which increases with the number of measurements one wants to make. Example:
   - Time: MRI.
   - Space: Imaging at micro-scale.
   - Material cost: Sensors components
2. Is taking many samples helpful in the first place? Example:
   - A 10MP digital photo has $\sim 10^7$ samples.
   - Will be compressed eventually (DCT, Wavelets): $\sim 10^5$ coefficients.

THE MRI EXAMPLE IN PARTICULAR

1. Fourier sampling in MRI angiography

$$\hat{f}(\omega_1, \omega_2) = \sum_{t_1, t_2} f(t_1, t_2) e^{-2\pi i(\omega_1 t_1 + \omega_2 t_2)}$$

2. The number of fourier coefficients sampled is very low compared to number of pixels desired in a final spatial image.
3. Takes a long time to get these.
4. For a 1MP image, 22 radial lines and 1000 samples on each line ($\sim$2% equations available to do inverse fourier!)
5. Using (filtered) Backprojection reconstruction doesn't yield good results.

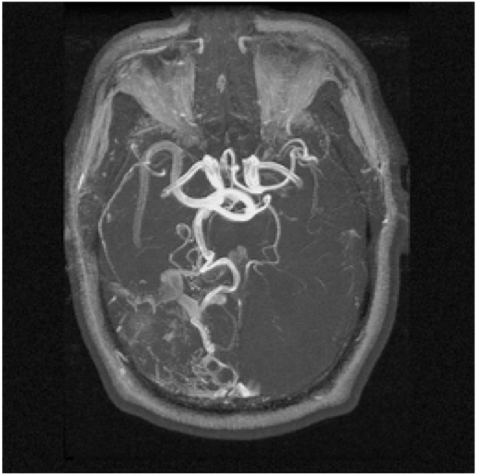


Figure: An MRI image.

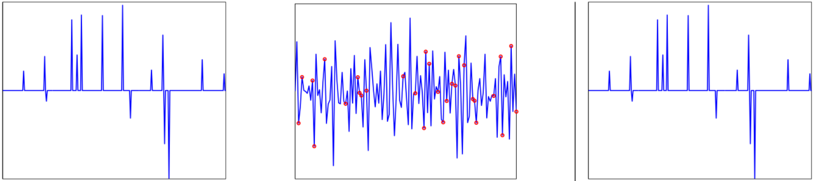
COMPRESSED SENSING FUNDAMENTALS

Compressive sampling theory says it is possible to undersample and expect perfect recovery.

The theory gives sampling schemes to extract information directly.

There are two important aspects of the theory: Sparsity and Incoherence.

But, some examples first.

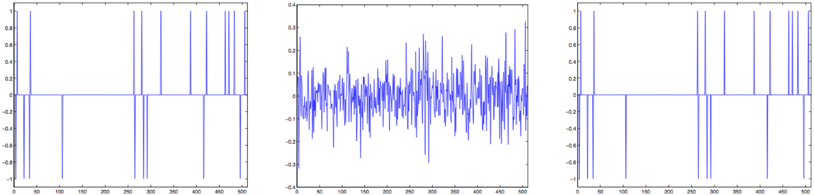


original $\hat{f}$, $S = 15$

given $m = 30$ time-dom. samples

perfect recovery

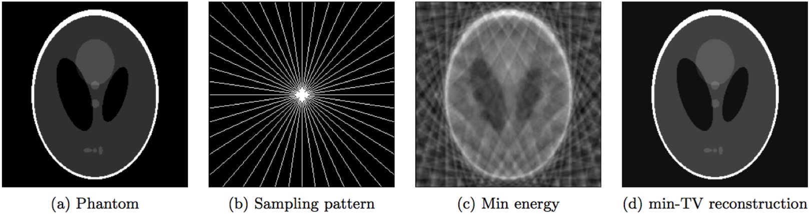
Figure: Fourier domain recovery from undersampled time signal.



(a) Original (b) Minimum energy reconstruction (c) Recovered

Figure 1: *1D recovery experiment for ℓ_1 minimization with equality constraints. (a) Original length 512 signal* **x** *consisting of 20 spikes. (b) Minimum energy (linear) reconstruction* **x0***. (c) Minimum ℓ_1 reconstruction* **xp***.*

Figure: CS reconstruction is exact with lesser samples than Nyquist thery dictates. ℓ_2 (or energy) minimization doesn't work.



(a) Phantom (b) Sampling pattern (c) Min energy (d) min-TV reconstruction

Figure 2: *Phantom recovery experiment.*

Figure: A synthetic example (relevant to MRI). TV means total variation. It is the sum of the magnitudes of discrete gradients at all points of an image. Gradient sparsity (sharp edges) is being exploited here.

SPARSITY AGAIN!

1. Here, let $f = \Phi x$. It is sparse in the basis Φ. That is, card(x) = s or x is s-sparse.
2. We can assume Φ is the canonical basis in R^n and just work with x being the sparse signal we want to reconstruct.
3. We have discussed sparsity at length.
4. If sparsity constraint is turned into a convex constraint, still reconstruction can be guaranteed.

PERFORMING COMPRESSIVE SAMPLING

1. Use a measurement matrix Φ. Sample m observations.
2. Let ℓ_0, ℓ_1 and ℓ_2 minimizations be:

$$\min ||x||_{\ell_0} \text{ s.t. } y =< \varphi_k, \Psi x >, k = 1, ..., m. \text{ (P0)}$$

$$\min ||x||_{\ell_1} \text{ s.t. } y =< \varphi_k, \Psi x >, k = 1, ..., m. \text{ (P1)}$$

$$\min ||x||_{\ell_2} \text{ s.t. } y =< \varphi_k, \Psi x >, k = 1, ..., m. \text{ (P2)}$$

3. Do ℓ_1 reconstruction. Success w.h.p if Restricted Isometry Property (RIP) condition is satisfied.

INCOHERENCE AND RIP

1. In Classicial, $\Phi = I_{n \times n}$ the identity matrix. We need n samples!
2. In CS, we are undersampling in a different basis.
3. The more uncorrelated the new basis (Φ) with the original one (Ψ), the better. Makes Φf dense.
4. Spread the information to the whole set of measurements.
5. This uncorrelatedness is ensured by RIP.
6. CS works because of the 'uncertainity principle': No signal can be sparse in two bases. (Donoho-Stark 1989). Positive result really.

ASIDE: UNCORRELATEDNESS

1. (Nazarov 1993) If $f \neq 0$ is concentrated on a set T $(1 - \epsilon$ fraction of its norm) and $\hat{f}$ is concentrated on a set Ω (again $1 - \delta$ fraction of its norm), then

$$|T||\Omega| > c \log \frac{c}{\epsilon + \delta}$$

2. Example: $f \in L_2(\mathbb{R})$and its fourier transform $\hat{f}$ cannot be both supported on compact sets.
3. **Discrete Uncertainty Principle** (Tao 2005) Assume n is prime. If $f \in \mathbb{R}^n$ is supported on a set T and $\hat{f}$ is supported on Ω, then

$$|T| + |\Omega| > n$$

4. Corollary: Assume n is prime. Then every s sparse signal $f \in \mathbb{R}^n$ is uniquely determined by the values of $\hat{f}$ at $2s$ points.

RIP STATEMENT

1. Reconstruction possible if the measurement matrix is bijective with the sparse vector x.
2. Algorithm exists w.h.p if Φ is an 'isometry' for $2s$ sparse or $3s$ sparse vectors.
3. (Candes, Tao 2004) Suppose the measurement matrix Φ is a restricted isometry, that is

$$(1-\delta_{4s})||x||_2 \leq ||\Phi x||_2 \leq (1+\delta_{4s})||x||_2$$

   then x can be reconstructed from measurements Φx as the solution to (P1).
4. RIP is nothing but uniform uncertainty principle for Φ.

Examples of Restricted Isometries

1. Since Φ is of size $m \times n$ it is not 1-1 on the whole space.
2. w.h.p, random matrices obey RIP.
   - $\Phi_{i,j} \sim N(0,1)$ then $m \sim s \log n$ (Candes, Tao, Rudelson, Vershynin 2005): Camera.
   - $\Phi_{i,j} \sim \{-1,+1\}$ w.p. $p = 0.5$ each, then $m \sim s \log n$ (Mendelson, Pajor, Tomczak 2006): Camera.
   - Random Fourier measurements: $m \sim s \log^4 n$ (Rudelson, Vershynin 2005): Medical Imaging.
3. Deterministic constructions: $m \sim sn^{o(1)}$ (Indyk 2007).

HARDWARE

1. MRI: Speedup factor of 6 (Mayo Clinic 2008)
2. Camera: 1-pixel camera by Kevin Kelly et al (RIce).

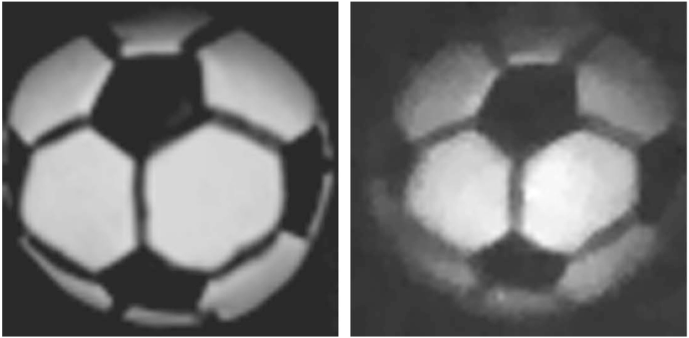


Figure: Figure of the football: 64x64 bit resolution (4096). Random measurements taken: 1600 (about 1/3$^{\text{rd}}$)

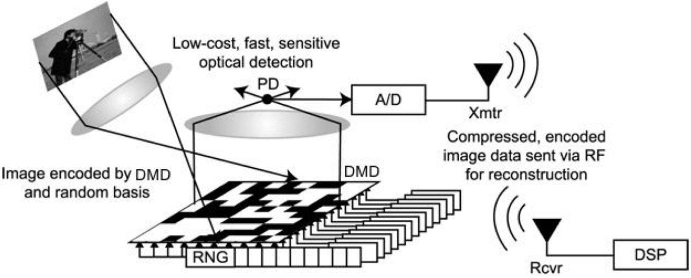

Low-cost, fast, sensitive optical detection
PD
A/D
Xmtr
Compressed, encoded image data sent via RF for reconstruction
Image encoded by DMD and random basis
DMD
RNG
Rcvr
DSP

Figure: An indicative diagram of the camera setup. The measurement matrix is binary (achieved by micro-mirrors).

RELATED TOPICS

1. Seismology
2. Duality with error correction coding
3. HIgh dimensional convex geometry
4. Compressible signals
5. Streaming algorithms: reconstruction in linear time w.r.t. sparsity

STREAMING: A COMPUTER SCIENCE PERSPECTIVE

1. Compressive sensing reverse the usual paradigm in source coding:
   - The encoder is resource poor. The decoder is resource rich.
   - Eg. Imaging, Sensor networks
2. Streaming algorithms go further:
   - Both encoder and decoder are resource poor.
   - Signals presented in streaming form or are too long to instantiate.
   - Eg: Inventory updates, Network data analysis

SUMMARY

1. Sparsity is a great heuristic. Are there others?
2. Compressive sampling: a technique to finding sparse solutions to underdetermined linear systems.
3. But caution: Not all underdetermined linear systems have a sparse solution.
4. Compressive Sampling works by:
   - Indirect undersampling (use of measurement matrix Φ)
   - Make use of incoherence and sparsity properties to reconstruct.