An introduction to compressed sensing

Ben Adcock
Department of Mathematics
Simon Fraser University

Overview

First talk (now): An introduction to compressed sensing.

- A (gentle) overview of the main principles of CS.

Second talk (later today): Compressed sensing and imaging.

Third talk (Monday): Compressed sensing and high-dimensional approximation.

The second and third talks will address two areas of application of compressed sensing. I will present more work of my own in these.

Outline

Outline

Basic problem

We wish to recover an unknown object from measurements.

- E.g. a signal, image, function, manifold,...

Finite-dimensional, linear setup

The object $x = (x_1, \ldots, x_N)^\top$ is a vector. The measurements are linear:

$$y = Ax + e, \qquad (\star)$$

where $A \in \mathbb{C}^{m \times N}$ is a measurement matrix and $e \in \mathbb{C}^N$ is noise.

Main issue: The number of measurements $m \ll N$, i.e. the system $(\star)$ is highly underdetermined.

Compressed sensing: the highlights

Under appropriate conditions on x and A we can recover x from y in a stable and robust manner. Moreover, this can be done with efficient numerical algorithms.

- Condition on x: low-dimensionality $s \ll N$.
- Condition on A: E.g. null space property, restricted isometry property, incoherence,...
- Condition on m: It is possible to find matrices A such that only $m \approx s \log(N/s)$ measurements suffice.
- Algorithms: convex optimization (ℓ^1 minimization), greedy methods, thresholding methods, message passing algorithms,...

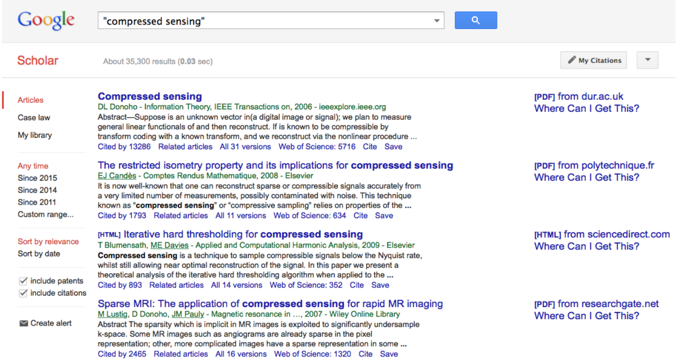
Compressed sensing: the highlights

Under appropriate conditions on x and A we can recover x from y in a stable and robust manner. Moreover, this can be done with efficient numerical algorithms.

- Condition on x: low-dimensionality $s \ll N$.
- Condition on A: E.g. null space property, restricted isometry property, incoherence,...
- Condition on m: It is possible to find matrices A such that only $m \approx s \log(N/s)$ measurements suffice.
- Algorithms: convex optimization (ℓ^1 minimization), greedy methods, thresholding methods, message passing algorithms,...

Compressed sensing: history

- Origins ($\approx$ 2004): Candès, Romberg & Tao, Donoho
- Since then, the subject of thousands of papers, dozens of survey articles, and one textbook (Foucart & Rauhut, Birkhauser, 2013).
- a.k.a. compressive sensing, compressed sampling, compressive sampling



Applications

Magnetic Resonance Imaging (MRI)

MRI uses magnetic fields to excite hydrogen atoms in the body, which result (after modelling) in measurements of the Fourier transform of the image. The number of measurements acquired is roughly proportional to the scan duration. Longer scans are unpleasant for patients, more prone to motion artifacts and impossible for certain parts of the body (e.g. those affected by breathing).

Image from the Wikipedia MRI page:
`http://en.wikipedia.org/wiki/Magnetic_resonance_imaging`

Applications

Compressive Imaging

Photodiodes in digital cameras are inexpensive an easily manufactured. For other types of imaging, e.g. infrared, the sensors are much more expensive and therefore one is limited to much lower resolutions with conventional imaging strategies. Compressive imaging exploits CS theory to recover images using only one sensor (single-pixel imaging) combined with a digital micromirror array.

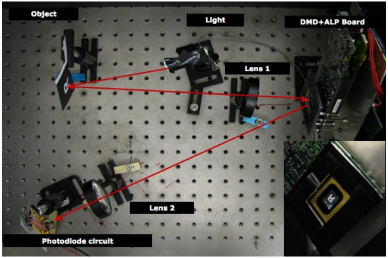


Image from Rice University: `http://dsp.rice.edu/cscamera`

Other applications

New applications are being investigated at a rapid rate. Some examples:

X-ray CT: Taking more measurements involves higher radiation doses.

Seismic Imaging: Sensors are placed on the surface of the earth to measure seismic waves. Limitations due to cost and geography.

Uncertainty Quantification: Goal is to predict how a physical system is affected by changes in parameters. Each measurement requires a time-consuming numerical solution of a PDE.

Also: electron microscopy, fluorescence microscopy, radar, analog-to-digital conversion,...

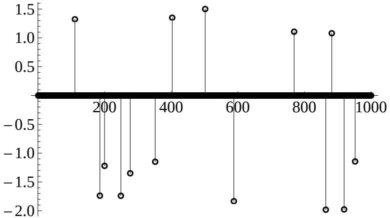
Outline

Sparsity

To recover $x \in \mathbb{C}^N$ from $y = Ax + e \in \mathbb{C}^m$ we require a low-dimensional model for x:

Definition (Sparsity)

A vector $x \in \mathbb{C}^N$ is *s*-sparse if it has at most s nonzero entries.



Remark: We may know s, but we do not know the locations on the nonzero entries.

Compressibility

Vectors are rarely exactly sparse, but they are often well approximated by sparse vectors.

Definition (Compressibility)

The best s-term approximation error is

$$\sigma_s(x) = \min \{ \|z - x\|_1 : z \in \Sigma_s \},$$

where Σ_s is the set of s-sparse vectors. A vector x is compressible if $\sigma_s(x)$ is small.

- Note that $\sigma_s(x) = \sum_{i=s+1}^{N} |x_i^*|$, where x^* is a rearrangement of x in nonincreasing order.
- That is, we approximate x by its largest s entries (hard thresholding).

Is sparsity a good model?

Canonical sparsity, i.e. the image/signal itself is sparse, is realistic in some applications, e.g. radar and astronomical imaging.

More typically, sparsity occurs in a transform domain:

- $x \in \mathbb{C}^N$ is the vector of coefficients of the unknown object

$$w = \Phi x \in \mathbb{C}^N,$$

with respect to some orthonormal basis $\Phi \in \mathbb{C}^{N \times N}$.

Examples:

- Fourier basis (sparse frequency estimation)
- DCT, wavelets, curvelets, shearlets (imaging)
- polynomials (high-dimensional approximation)
- ...

Images are compressible in wavelets

This principle underlies many modern lossy compression formats, e.g. JPEG-2000. Since $x = \Phi^* w$ is approximately sparse, one only needs to store a small fraction of its entries.

Image w

Images are compressible in wavelets

This principle underlies many modern lossy compression formats, e.g. JPEG-2000. Since $x = \Phi^* w$ is approximately sparse, one only needs to store a small fraction of its entries.

Image $\tilde{w}$ obtained by keeping only the largest 4% wavelet coefficients

Images are compressible in wavelets

This principle underlies many modern lossy compression formats, e.g. JPEG-2000. Since $x = \Phi^* w$ is approximately sparse, one only needs to store a small fraction of its entries.

Image $\tilde{w}$ obtained by keeping only the largest 0.5% wavelet coefficients

Beyond sparsity

Sparsity is only a model and may not be the best fit for all applications.

Some extensions:

1. Model/tree sparsity
   - Baraniuk et al. (2010), Carin et al. (2009), Schniter et al. (2010),...
   - Wavelet coefficients lie on connected trees
2. Block sparsity
   - Eldar et al. (2010), Lu & Do (2008), Blumensath & Davies (2009),...
   - Nonzero coefficients cluster in unknown blocks
3. Joint sparsity
   - Duarte et al. (2009), Fornasier & Rauhut (2008),...
   - Collections of vectors with common support sets.

Outline

Recovery of sparse vectors

The following lemma points towards a recovery algorithm:

Lemma

Let $A \in \mathbb{C}^{m \times N}$ and let $x \in \mathbb{C}^N$ be s-sparse. The following two statements are equivalent:

(i) *x is the unique s-sparse solution of $Az = y$, where $y = Ax$.*

(ii) *x is the unique solution to*

$$\min_{z \in \mathbb{C}^N} \|z\|_0 \text{ subject to } Az = y, \qquad (\star)$$

where $\|z\|_0 = |\{j : z_j \neq 0\}|$.

Unfortunately $(\star)$ is NP-hard to solve in general, and therefore impractical for computations.

Convex relaxation

To obtain a computationally tractable problem, we make a convex relaxation. We replace

$$\min_{z \in \mathbb{C}^N} \|z\|_0 \ \text{ subject to } Az = y,$$

by

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \ \text{ subject to } Az = y, \qquad (\star)$$

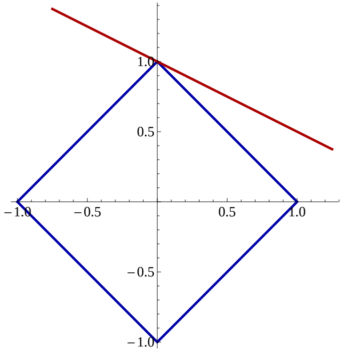
where $\|\cdot\|_1$ is the l^1-norm.

Many algorithms exist for solving the convex problem $(\star)$. E.g.

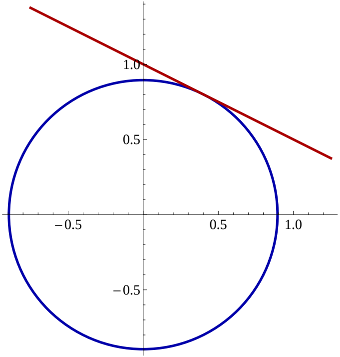
- homotopy methods, LARS, primal dual algorithms, pareto curve methods, iteratively reweighted least squares, splitting methods (e.g. split Bregman), ADMM,...

Why l^1 and why not l^2?

Consider the case $N = 2$, $m = 1$ and $s = 1$. The possible solutions lie on the line $L(y) = x + \ker(A)$.



l^1 solution tends to be sparse



l^2 solution tends not to be sparse

The Restricted Isometry Property

One of the most popular ways to analyze the recovery performance of l^1 minimization and other CS algorithms.

Definition

The restricted isometry constant δ_s of a matrix $A \in \mathbb{C}^{m \times N}$ is the smallest number such that

$$(1 - \delta_s)\|z\|_2^2 \leq \|Az\|_2^2 \leq (1 + \delta_s)\|z\|_2^2, \qquad \forall z \in \Sigma_s.$$

We say A satisfies the Restricted Isometry Property (RIP) of order s with constant δ_s if $\delta_s \in (0, 1)$.

Alternatives:

- (Robust) Null Space Property
- Coherence (version 1) – however, suffers from a quadratic bottleneck
- Coherence (version 2) – see later

Explanation

Suppose we knew the support set of x, i.e.

$$\Delta = \{j : x_j \neq 0\} \subseteq \{1, \dots, N\}.$$

Let A_Δ be the matrix formed by the columns of A with indices in Δ and consider the overdetermined $m \times s$ system:

$$A_\Delta x = y.$$

Under the condition

$$\delta_\Delta = \|A_\Delta^* A_\Delta - I_\Delta\|_2 \in (0, 1),$$

the matrix $A_\Delta^* A_\Delta$ is nonsingular and well-conditioned, and we can recover x stably and robustly by

$$x = A_\Delta^\dagger y.$$

Explanation continued

The RIC δ_s, defined by

$$(1-\delta_s)\|z\|_2^2 \leq \|Az\|_2^2 \leq (1+\delta_s)\|z\|_2^2, \qquad \forall z \in \Sigma_s,$$

is equivalent to

$$\delta_s = \max_{\substack{\Delta \subseteq \{1,\ldots,N\} \\ |\Delta| \leq s}} \|A_\Delta^* A_\Delta - I_\Delta\|_2.$$

Hence the RIP ensures that the oracle method, based on knowing Δ, works for any s-sparse vector x.

Stable and robust recovery with the RIP

Theorem

Suppose that the matrix $A \in \mathbb{C}^{m \times N}$ satisfies the RIP of order $2s$ with constant

$$\delta_{2s} < 4/\sqrt{41}. \qquad (\star)$$

Then for any $x \in \mathbb{C}^N$ and $y \in \mathbb{C}^m$ with $\|Ax - y\|_2 \leq \epsilon$, any solution $\hat{x}$ of

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \text{ subject to } \|Az - y\|_2 \leq \epsilon,$$

satisfies

$$\|x - \hat{x}\|_2 \leq C_1 \frac{\sigma_s(x)}{\sqrt{s}} + C_2 \epsilon.$$

Note: There are many variants of $(\star)$.

Remark: The RIP also ensures recovery for other algorithms for CS (e.g. greedy or thresholding methods).

Matrices that satisfy the RIP

Deterministic construction of RIP matrices with m scaling linearly with s have proved elusive. This is perhaps understandable, since the condition

$$\delta_s = \max_{\substack{\Delta \subseteq \{1,\ldots,N\} \\ |\Delta| \leq s}} \|A_\Delta^* A_\Delta - I_\Delta\|_2 \in (0,1),$$

is inherently combinatorial.

- Moreover, future deterministic constructions, if at all possible, are likely to be impractical (i.e. memory/CPU intensive).

The major breakthrough in CS was to consider random constructions.

Random matrices satisying the RIP

Theorem (Candès & Tao, Mendelson et al., Baraniuk et al.)

Suppose that $A \in \mathbb{C}^{m\times N}$ is a Gaussian or Bernoulli random matrix. Then, with probability at least $1-\epsilon$, the matrix $\frac{1}{\sqrt{m}}A$ satisfies the RIP with constant $\delta_s \leq \delta$, provided

$$m \geq C\delta^{-2}\left(s\log(\mathrm{e}N/s)+\log(2\epsilon^{-1})\right).$$

- This result is fundamental. However, such matrices are largely computationally infeasible.

Theorem (Rudelson & Vershynin, Rauhut, Andersson & Strömberg)

Let $A \in \mathbb{C}^{m\times N}$ be formed by drawing m rows of the Fourier matrix $F \in \mathbb{C}^{N\times N}$ uniformly at random. Then, with probability at least $1-\epsilon$, the matrix $\sqrt{\frac{N}{m}}A$ satisfies the RIP with constant $\delta_s \leq \delta$, provided

$$m \geq C\delta^{-2}s\left(\log^3(s)\log(N)+\log(\epsilon^{-1})\right).$$

Random matrices satisying the RIP

Theorem (Candès & Tao, Mendelson et al., Baraniuk et al.)

Suppose that $A \in \mathbb{C}^{m \times N}$ is a Gaussian or Bernoulli random matrix. Then, with probability at least $1-\epsilon$, the matrix $\frac{1}{\sqrt{m}}A$ satisfies the RIP with constant $\delta_s \leq \delta$, provided

$$m \geq C\delta^{-2}\left(s\log(\mathrm{e}N/s) + \log(2\epsilon^{-1})\right).$$

- This result is fundamental. However, such matrices are largely computationally infeasible.

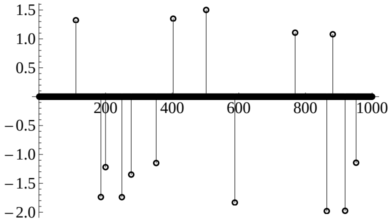
Theorem (Rudelson & Vershynin, Rauhut, Andersson & Strömberg)

Let $A \in \mathbb{C}^{m \times N}$ be formed by drawing m rows of the Fourier matrix $F \in \mathbb{C}^{N \times N}$ uniformly at random. Then, with probability at least $1-\epsilon$, the matrix $\sqrt{\frac{N}{m}}A$ satisfies the RIP with constant $\delta_s \leq \delta$, provided

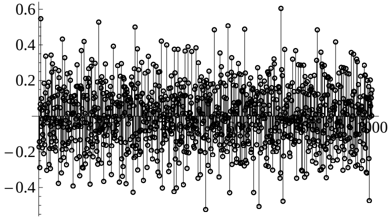
$$m \geq C\delta^{-2}s\left(\log^3(s)\log(N) + \log(\epsilon^{-1})\right).$$

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



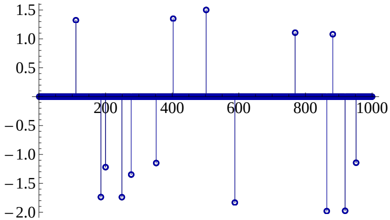
x



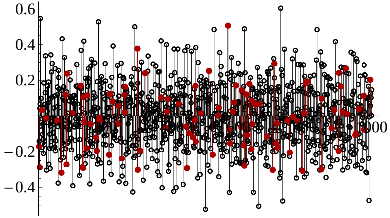
Fx

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



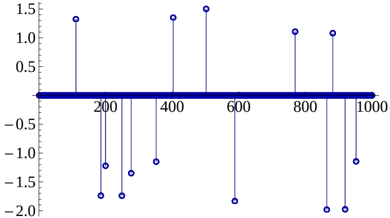
$\hat{x}$ with $m = 100$



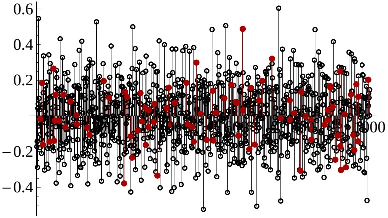
Ax

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



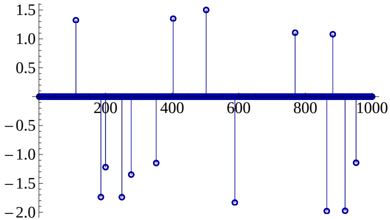
$\hat{x}$ with $m = 90$



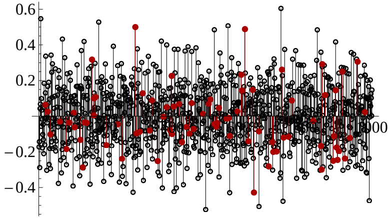
Ax

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



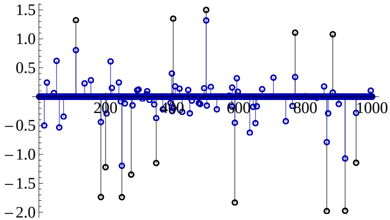
$\hat{x}$ with $m = 80$



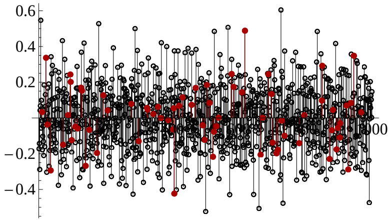
Ax

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



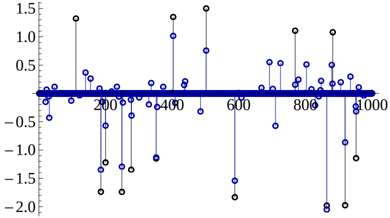
$\hat{x}$ with $m = 70$



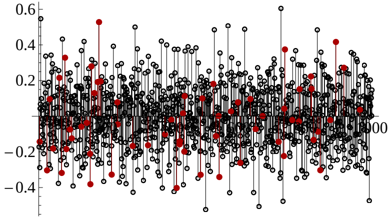
Ax

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



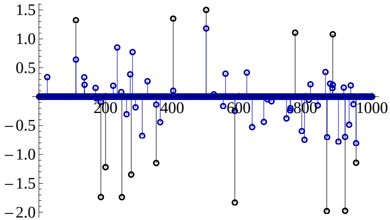
$\hat{x}$ with $m = 60$



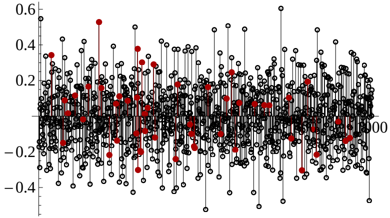
Ax

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



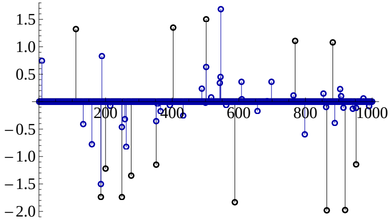
$\hat{x}$ with $m = 50$



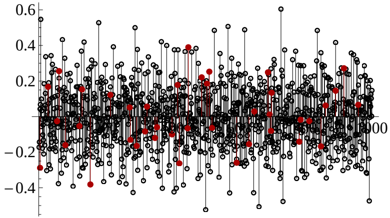
Ax

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



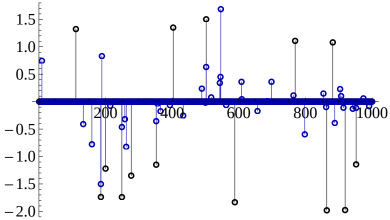
$\hat{x}$ with $m = 40$



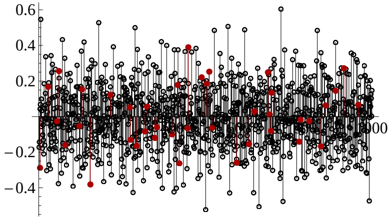
Ax

Random subsampling the Fourier matrix

An example with $N = 1000$ and $s = 20$:



$\hat{x}$ with $m = 40$



Ax

Phase transition behaviour is typical in CS.

Outline

Problems with the RIP

1. The RIP is NP-hard to verify in general.

2. The RIP often leads to a more stringent measurement condition (e.g. additional log factors).

Explanation: the RIP stipulates uniform recovery.

- One draw of the matrix (e.g. subsampled Fourier) is sufficient to recover all s-sparse vectors, with high probability.

This may be too pessimistic in practice. Instead, we can consider the substantially weaker notion of nonuniform recovery:

- One draw of the matrix is sufficient to recover a fixed s-sparse vector, with high probability.

Problems with the RIP

1. The RIP is NP-hard to verify in general.

2. The RIP often leads to a more stringent measurement condition (e.g. additional log factors).

Explanation: the RIP stipulates uniform recovery.

- One draw of the matrix (e.g. subsampled Fourier) is sufficient to recover all s-sparse vectors, with high probability.

This may be too pessimistic in practice. Instead, we can consider the substantially weaker notion of nonuniform recovery:

- One draw of the matrix is sufficient to recover a fixed s-sparse vector, with high probability.

Subsampled isometries

Setup:

- Let $U = (u_{ij}) \in \mathbb{C}^{N \times N}$ be an isometry, e.g. the Fourier matrix.
- We form $A \in \mathbb{C}^{m \times N}$ by drawing m rows of U uniformly at random.

Remark:

- This setup is less general than before.
- It can be generalized somewhat to include, for example, random Gaussian matrices (see Candès & Plan).

Incoherence

Definition

The coherence of U is

$$\mu(U) = \max_{i,j} |u_{ij}|^2 \in [N^{-1}, 1].$$

U is incoherent if $\mu(U) \leq c/N$ for some $c \geq 1$ independent of N.

- E.g. For the Fourier matrix $F = U$, we have $\mu(U) = 1/N$.

Main claim: If U is incoherent, then we can recover any s-sparse vector from $m \approx s \cdot \log(\epsilon^{-1}) \cdot \log(N)$ rows of U chosen uniformly at random.

Remarks:

- Incoherence is much easier to check than the RIP.
- Fewer log factors.

Incoherence

Definition

The coherence of U is

$$\mu(U) = \max_{i,j} |u_{ij}|^2 \in [N^{-1}, 1].$$

U is incoherent if $\mu(U) \leq c/N$ for some $c \geq 1$ independent of N.

- E.g. For the Fourier matrix $F = U$, we have $\mu(U) = 1/N$.

Main claim: If U is incoherent, then we can recover any s-sparse vector from $m \approx s \cdot \log(\epsilon^{-1}) \cdot \log(N)$ rows of U chosen uniformly at random.

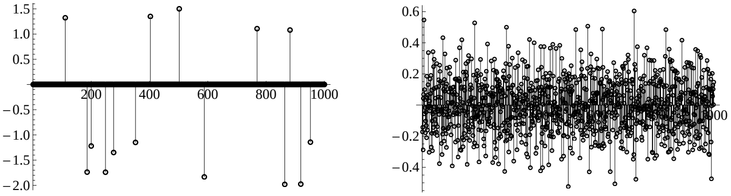
Remarks:

- Incoherence is much easier to check than the RIP.
- Fewer log factors.

Intuition

Incoherence relates to a discrete uncertainty principle.

- If x is sparse, then Ux cannot be sparse.
- Specifically, $\|x\|_0 + \|Ux\|_0 \geq 2/\sqrt{\mu(U)}$.
- See Donoho & Starck, Elad & Bruckstein,...



x $\qquad$ Ux

Since the information about x is spread across the entries of Ux, random sampling should be sufficient to guarantee recovery of x from $y = Ax$.

Main theorem

Theorem (see Candès & Plan, BA & Hansen)

Let $0 < \epsilon \leq \mathrm{e}^{-1}$ *and suppose that*

$$m \geq C \cdot s \cdot N \cdot \mu(U) \cdot \log(\epsilon^{-1}) \cdot \log N.$$

Then with probability greater than $1 - \epsilon$ *any minimizer* $\hat{x}$ *of the problem*

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \text{ subject to } \|Az - y\|_2 \leq \delta \sqrt{N/m},$$

satisfies

$$\|x - \hat{x}\|_{l^2} \leq C_1 \sigma_s(x) + C_2 L \sqrt{s} \delta,$$

where $L = 1 + \frac{\sqrt{\log(\epsilon^{-1})}}{\log(4N\sqrt{s})}$.

If U is incoherent, i.e. $\mu(U) \lesssim 1/N$, then we get $m \approx s \cdot \log(\epsilon^{-1}) \cdot \log(N)$.

Ideas behind the proof

The proof is based on constructing a certain dual certificate:

Lemma

Let $\Delta \subseteq \{1,\dots,N\}$, $|\Delta| = s$ be the support of the largest s entries of x. Suppose that A is such that

(i) $\|P_\Delta A^* A P_\Delta - P_\Delta\|_2 \leq \alpha$,

(ii) $\max_{i \notin \Delta} \{\|Ae_i\|_2\} \leq \beta$,

and that there exists a vector $\rho = A^*\xi$ for some $\xi \in \mathbb{C}^m$ such that

(iii) $\|W(P_\Delta \rho - \mathrm{sign}(P_\Delta x))\|_2 \leq \gamma$,

(iv) $\|P_\Delta^\perp \rho\|_\infty \leq \theta$,

(v) $\|\xi\|_2 \leq \lambda\sqrt{|\Delta|_w}$,

for $0 \leq \alpha, \theta < 1$ and $\beta, \gamma, \lambda \geq 0$ satisfying $\frac{\sqrt{1+\alpha}\beta\gamma}{(1-\alpha)(1-\theta)} < 1$. Then the conclusions of the theorem hold with $L = \lambda$ and appropriate C_1 and C_2.

Note: (i) and (ii) can be verified using the (matrix) Bernstein inequality.

Constructing the dual certificate

The construction of the dual certificate ρ uses an iterative construction known as the golfing scheme and due to D. Gross.

- First, one divides the rows of A into L bins, of sizes $m_1, \dots, m_L$.
- Set $\rho^{(0)} = 0$.
- For $l = 1, \dots, L$ perform the iterative update

$$\rho^{(l)} = m_l^{-1}(A^{(l)})^* A^{(l)} \left(\mathrm{sign}(P_\Delta x) - P_\Delta \rho^{(l-1)}\right) + \rho^{(l-1)},$$

  provided
  - $\|(P_\Delta - m_l^{-1} P_\Delta (A^{(l)})^* A^{(l)} P_\Delta) v^{(l-1)}\|_2 \leq a_l \|v^{(l-1)}\|_2$,
  - $\|m^{-1} P_\Delta^\perp (A^{(l)})^* A^{(l)} P_\Delta v^{(l-1)}\|_\infty \leq b_l \|v^{(l-1)}\|_2$,

  where $v^{(l)} = \mathrm{sign}(P_\Delta x) - P_\Delta \rho^{(l)}$.
- The parameters $m_1, \dots, m_L$, L, a_l, b_l are carefully tuned to get the correct recovery guarantee.

Outline

Conclusions

1. Compressed sensing concerns the recovery of sparse vectors from limited measurements.

2. Recovery can be achieved via ℓ^1 minimization, although other techniques are possible.

3. The RIP provides a sufficient condition for recovery. However, it is hard to verify and may be too stringent.

4. A more intuitive and easier-to-verify condition is provided incoherence. This dictates that measurements must 'spread out' information.

Topics not covered:

- Other algorithms: greedy, thresholding methods, nonconvex optimization, reweighted ℓ^1 minimization.
- Empirical recovery performance via phase transitions.
- Other notions of coherence.
- Redundant sparsifying transforms, e.g. TV minimization.

Conclusions

1. Compressed sensing concerns the recovery of sparse vectors from limited measurements.

2. Recovery can be achieved via ℓ^1 minimization, although other techniques are possible.

3. The RIP provides a sufficient condition for recovery. However, it is hard to verify and may be too stringent.

4. A more intuitive and easier-to-verify condition is provided incoherence. This dictates that measurements must 'spread out' information.

Topics not covered:

- Other algorithms: greedy, thresholding methods, nonconvex optimization, reweighted ℓ^1 minimization.
- Empirical recovery performance via phase transitions.
- Other notions of coherence.
- Redundant sparsifying transforms, e.g. TV minimization.

Outline

Recovery from the Fourier transform

Applications: Magnetic Resonance Imaging (MRI), X-ray Computed Tomography, Electron Microscopy, Seismic Imaging, Radio interferometry,....

Mathematically, all these problems can be reduced (possibly via the Fourier slice theorem) to the following:

$$\text{Given } \{\hat{f}(\omega) : \omega \in \Omega\}, \text{ recover the image } f.$$

Here $\Omega \subseteq \hat{\mathbb{R}}^d$ is a finite set and $\hat{f}$ is the Fourier transform (FT).

However, f is a function (not a vector) and $\hat{f}$ is its continuous FT.

Standard CS approach

We approximate $f \approx x$ on a discrete grid, and let

- F be the Fourier matrix,
- Φ be a discrete wavelet transform,
- and set $U = F^* \Phi$

However, this setup is a discretization of the continuous model:

$$\text{continuous FT} \approx \text{discrete FT} \quad \Rightarrow \quad \text{measurements mismatch}$$

Issues:

1. If measurements are simulated via the DFT $\Rightarrow$ inverse crime.
   - In MRI, see Guerquin–Kern, Häberlin, Pruessmann & Unser (2012)
2. If measurements are simulated via the continuous FT. Minimization problem has no sparse solution $\Rightarrow$ poor reconstructions.

Standard CS approach

We approximate $f \approx x$ on a discrete grid, and let

- F be the Fourier matrix,
- Φ be a discrete wavelet transform,
- and set $U = F^* \Phi$

However, this setup is a discretization of the continuous model:

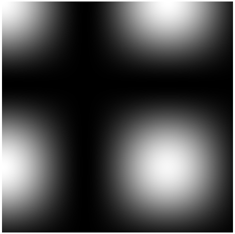
continuous FT $\approx$ discrete FT $\Rightarrow$ measurements mismatch

Issues:

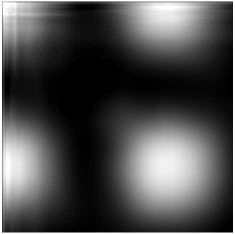
1. If measurements are simulated via the DFT $\Rightarrow$ inverse crime.
   - In MRI, see Guerquin–Kern, Häberlin, Pruessmann & Unser (2012)
2. If measurements are simulated via the continuous FT. Minimization problem has no sparse solution $\Rightarrow$ poor reconstructions.

Poor reconstructions with standard CS

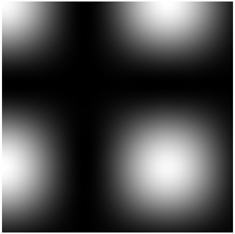
Example: Electron microscopy, $f(x,y) = \mathrm{e}^{-x-y}\cos^2(\frac{17\pi x}{2})\cos^2(\frac{17\pi y}{2})$, 6.15% Fourier measurements.



Original (zoomed)



Fin. dim. CS, Err = 12.7%



Inf. dim. CS, Err = 0.6%

Infinite-dimensional setup

Consider two orthonormal bases of a Hilbert space H (e.g. $\mathrm{L}^2(0,1)^d$):

- Sampling basis: $\{\psi_j\}_{j\in\mathbb{N}}$, e.g. the Fourier basis $\psi_j(x) = \exp(2\pi \mathrm{i} j \cdot x)$.
- Sparsity basis: $\{\phi_j\}_{j\in\mathbb{N}}$, e.g. a wavelet basis.

Let $f \in \mathrm{H}$ be the object to recover. Write

- $x_j = \langle f, \phi_j \rangle$ for the coefficients of f, i.e. $f = \sum_{j\in\mathbb{N}} x_j \phi_j$,
- $y_j = \langle f, \psi_j \rangle$ for the measurements of f.

Define the infinite matrix $U = \{\langle \phi_j, \psi_i \rangle\}_{i,j\in\mathbb{N}} \in B(\ell^2(\mathbb{N}))$ and note that

$$Ux = y.$$

As in finite dimensions, operator U is an isometry.

- BA & Hansen, *Generalized sampling and infinite-dimensional compressed sensing*, Found. Comput. Math. (to appear), 2015.

New concepts

To generalize CS to this Hilbert space setting, we need analogues of the key concepts:

- Sparsity
- Uniform random subsampling
- Incoherence

New concepts

Uniform random subsampling: It is meaningless to draw $\Omega \subseteq \mathbb{N}$, $|\Omega| = m$ uniformly at random. It is also infeasible in practice due to bandwidth limitations. Hence, we fix the sampling bandwidth N and let

$$\Omega \subseteq \{1, \ldots, N\}, \quad |\Omega| = m,$$

be drawn uniformly at random.

Sparsity: Given finite sampling bandwidth, we cannot expect to recover any s-sparse infinite vector x stably. Let M be the sparsity bandwidth, and suppose that x is (s, M)-sparse:

$$|\{j = 1, \ldots, M : x_j \neq 0\}| \leq s, \qquad x_j = 0, \ j > M.$$

Coherence: Define $\mu(U) = \sup |u_{ij}|^2$ as before.

New concepts

Uniform random subsampling: It is meaningless to draw $\Omega \subseteq \mathbb{N}$, $|\Omega| = m$ uniformly at random. It is also infeasible in practice due to bandwidth limitations. Hence, we fix the sampling bandwidth N and let

$$\Omega \subseteq \{1, \ldots, N\}, \quad |\Omega| = m,$$

be drawn uniformly at random.

Sparsity: Given finite sampling bandwidth, we cannot expect to recover any s-sparse infinite vector x stably. Let M be the sparsity bandwidth, and suppose that x is (s, M)-sparse:

$$|\{j = 1, \ldots, M : x_j \neq 0\}| \leq s, \qquad x_j = 0, \; j > M.$$

Coherence: Define $\mu(U) = \sup |u_{ij}|^2$ as before.

New concepts

Uniform random subsampling: It is meaningless to draw $\Omega \subseteq \mathbb{N}$, $|\Omega| = m$ uniformly at random. It is also infeasible in practice due to bandwidth limitations. Hence, we fix the sampling bandwidth N and let

$$\Omega \subseteq \{1, \dots, N\}, \quad |\Omega| = m,$$

be drawn uniformly at random.

Sparsity: Given finite sampling bandwidth, we cannot expect to recover any s-sparse infinite vector x stably. Let M be the sparsity bandwidth, and suppose that x is (s, M)-sparse:

$$|\{j = 1, \dots, M : x_j \neq 0\}| \leq s, \qquad x_j = 0, \ j > M.$$

Coherence: Define $\mu(U) = \sup |u_{ij}|^2$ as before.

The balancing property

Let $P_N : l^2(\mathbb{N}) \to l^2(\mathbb{N})$ be the projection onto the first N elements, i.e. $P_N x = \{x_1, \ldots, x_N, 0, 0, \ldots\}$, $x \in l^2(\mathbb{N})$.

Key idea: Given a sparsity bandwidth M, we need to take the sampling bandwidth N sufficiently large.

Definition (The balancing property)

$N \in \mathbb{N}$ and $K \geq 1$ satisfy the strong balancing property with respect to $s, M \in \mathbb{N}$ if

(i) $\|P_M U^* P_N U P_M - P_M\|_{l^\infty} \leq \frac{1}{8}\left(\log_2(4\sqrt{s}KM)\right)^{-1/2}$,

(ii) $\|P_M^\perp U^* P_N U P_M\|_{l^\infty} \leq \frac{1}{8}$.

The balancing property

Let $P_N : l^2(\mathbb{N}) \to l^2(\mathbb{N})$ be the projection onto the first N elements, i.e. $P_N x = \{x_1, \ldots, x_N, 0, 0, \ldots\}$, $x \in l^2(\mathbb{N})$.

Key idea: Given a sparsity bandwidth M, we need to take the sampling bandwidth N sufficiently large.

Definition (The balancing property)

$N \in \mathbb{N}$ and $K \geq 1$ satisfy the strong balancing property with respect to $s, M \in \mathbb{N}$ if

(i) $\|P_M U^* P_N U P_M - P_M\|_{l^\infty} \leq \frac{1}{8} \left(\log_2(4\sqrt{s}KM)\right)^{-1/2}$,

(ii) $\|P_M^\perp U^* P_N U P_M\|_{l^\infty} \leq \frac{1}{8}$.

Explanation

As we have seen, isometries are highly desirable in CS.

- The infinite matrix U is an isometry.
- However, we cannot work with U since it is infinite.
- The balancing property ensures that the uneven section $P_N U P_M$, a finite matrix, is close to an isometry.
- In particular,

$$\lim_{N\to\infty} \|P_M U^* P_N U P_M - P_M\|_{l^\infty} = \lim_{N\to\infty} \|P_M^\perp U^* P_N U P_M\|_{l^\infty} = 0.$$

Note: Typically we cannot take $M = N$, i.e. the finite section $P_N U P_N$.

Examples: For Fourier/wavelets, we need $N = \mathcal{O}(M)$. For Fourier/polynomials, we need $N = \mathcal{O}\left(M^2\right)$.

Infinite-dimensional CS theorem

Theorem (BA & Hansen)

Suppose that $N \in \mathbb{N}$ *and* $K = N/m \geq 1$ *satisfy the strong balancing property with respect to* $s, M \in \mathbb{N}$ *and also, for some for* $0 < \epsilon \leq \mathrm{e}^{-1}$,

$$m \gtrsim s \cdot N \cdot \mu(U) \cdot \log(\epsilon^{-1}) \cdot \log(K \tilde{M} \sqrt{s})$$

where $\tilde{M} = \min\{i \in \mathbb{N} : \max_{k \geq i} \|P_N U P_{\{i\}}\| \leq 1/(32K\sqrt{s})\}$. *If* $\hat{x}$ *is any minimizer of*

$$\inf_{z \in l^1(\mathbb{N})} \|z\|_{l^1} \text{ subject to } \|P_\Omega U z - y\|_{l^2} \leq \delta\sqrt{K},$$

then

$$\|x - \hat{x}\|_{l^2} \lesssim \sigma_{s,M}(x) + L\sqrt{s}\delta,$$

where $\sigma_{s,M}(x) = \min\{\|x - z\|_{l^1} : z \in \Sigma_{s,M}\}$ *and* $L = 1 + \frac{\sqrt{\log(\epsilon^{-1})}}{\log(4KM\sqrt{s})}$.

- BA & Hansen, *Generalized sampling and infinite-dimensional compressed sensing*, Found. Comput. Math. (to appear), 2015.

Remarks

Ideas of the proof:

- Relate the problem to that of constructing a dual certificate.
- Use a golfing-type scheme adapted to infinite dimensions.
- Main technical issues are handling the infinite-dimensionality of the problem, i.e. estimating tails using the balancing property.

Remark:

- Finite-dimensional, incoherence-based CS becomes a direct corollary of this result.

Open problem:

- How do we prove uniform recovery results in infinite dimensions.
- No known generalization of the RIP to this setting.

Remarks

Ideas of the proof:

- Relate the problem to that of constructing a dual certificate.
- Use a golfing-type scheme adapted to infinite dimensions.
- Main technical issues are handling the infinite-dimensionality of the problem, i.e. estimating tails using the balancing property.

Remark:

- Finite-dimensional, incoherence-based CS becomes a direct corollary of this result.

Open problem:

- How do we prove uniform recovery results in infinite dimensions.
- No known generalization of the RIP to this setting.