Sparse Optimization
Lecture: Sparse Recovery Guarantees

Instructor: Wotao Yin

July 2013

Note scriber: Zheng Sun
online discussions on piazza.com

Those who complete this lecture will know

- how to read different recovery guarantees
- some well-known conditions for exact and stable recovery such as Spark, coherence, RIP, NSP, etc.
- indeed, we can trust ℓ_1-minimization for recovering sparse vectors

The basic question of sparse optimization is:

Can I trust my model to return an intended sparse quantity?

That is

- does my model have a unique solution? (otherwise, different algorithms may return different answers)
- is the solution exactly equal to the original sparse quantity?
- if not (due to noise), is the solution a faithful approximate of it?
- how much effort is needed to numerically solve the model?

This lecture provides brief answers to the first three questions.

What this lecture does and does not cover

It **covers** basic sparse vector recovery guarantees based on

- spark
- coherence
- restricted isometry property (RIP) and null-space property (NSP)

as well as both exact and robust recovery guarantees.

It does **not cover** the recovery of matrices, subspaces, etc.

Recovery guarantees are important parts of sparse optimization, but they are *not* the focus of this summer course.

Examples of guarantees

Theorem (Donoho and Elad [2003], Gribonval and Nielsen [2003])

For $\mathbf{A}\mathbf{x} = \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$ has full rank, if $\mathbf{x}$ satisfies $\|\mathbf{x}\|_0 \leq \frac{1}{2}(1 + \mu(\mathbf{A})^{-1})$, then ℓ_1-minimization recovers this $\mathbf{x}$.

Theorem (Candes and Tao [2005])

If $\mathbf{x}$ is k-sparse and $\mathbf{A}$ satisfies the RIP-based condition $\delta_{2k} + \delta_{3k} < 1$, then $\mathbf{x}$ is the ℓ_1-minimizer.

Theorem (Zhang [2008])

IF $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a standard Gaussian matrix, then with probability at least $1 - \exp(-c_0(n-m))$ ℓ_1-minimization is equivalent to ℓ_0-minimization for all $\mathbf{x}$:

$$\|\mathbf{x}\|_0 < \frac{c_1^2}{4} \frac{m}{1 + \log(n/m)}$$

where $c_0, c_1 > 0$ are constants independent of m and n.

How to read guarantees

Some basic aspects that distinguish different types of guarantees:

- Recoverability (exact) vs stability (inexact)
- General $\mathbf{A}$ or special $\mathbf{A}$?
- Universal (all sparse vectors) or instance (certain sparse vector(s))?
- General optimality? or specific to model / algorithm?
- Required property of $\mathbf{A}$: spark, RIP, coherence, NSP, dual certificate?
- If randomness is involved, what is its role?
- Condition/bound is tight or not? Absolute or in order of magnitude?

Spark

First questions for finding the sparsest solution to $\mathbf{Ax} = \mathbf{b}$

1. Can sparsest solution be unique? Under what conditions?
2. Given a sparse $\mathbf{x}$, how to verify whether it is actually the sparsest one?

Definition (Donoho and Elad [2003])

The *spark* of a given matrix $\mathbf{A}$ is the smallest number of columns from $\mathbf{A}$ that are linearly dependent, written as $\mathrm{spark}(\mathbf{A})$.

$\mathrm{rank}(\mathbf{A})$ is the largest number of columns from $\mathbf{A}$ that are linearly independent. In general, $\mathrm{spark}(\mathbf{A}) \neq \mathrm{rank}(\mathbf{A}) + 1$; except for many randomly generated matrices.

Rank is easy to compute (due to the *matroid* structure), but spark needs a combinatorial search.

Spark

Theorem (Gorodnitsky and Rao [1997])

If $\mathbf{Ax} = \mathbf{b}$ has a solution $\mathbf{x}$ obeying $\|\mathbf{x}\|_0 < \mathrm{spark}(\mathbf{A})/2$, then $\mathbf{x}$ is the sparsest solution.

- **Proof idea**: if there is a solution $\mathbf{y}$ to $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{x} - \mathbf{y} \neq 0$, then $\mathbf{A}(\mathbf{x} - \mathbf{y}) = 0$ and thus

$$\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \|\mathbf{x} - \mathbf{y}\|_0 \geq \mathrm{spark}(\mathbf{A})$$

  or $\|\mathbf{y}\|_0 \geq \mathrm{spark}(\mathbf{A}) - \|\mathbf{x}\|_0 > \mathrm{spark}(\mathbf{A})/2 > \|\mathbf{x}\|_0$.

- The result does not mean this $\mathbf{x}$ can be efficiently found numerically.
- For many random matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, the result means that if an algorithm returns $\mathbf{x}$ satisfying $\|\mathbf{x}\|_0 < (m+1)/2$, the $\mathbf{x}$ is optimal with probability 1.
- What to do when $\mathrm{spark}(\mathbf{A})$ is difficult to obtain?

General Recovery - Spark

Rank is easy to compute, but spark needs a combinatorial search.

However, for matrix with entries in general positions, $\text{spark}(\mathbf{A}) = \text{rank}(\mathbf{A}) + 1$.

For example, if matrix $\mathbf{A} \in \mathbb{R}^{m \times n} (m < n)$ has entries $A_{ij} \sim \mathcal{N}(0, 1)$, then $\text{rank}(\mathbf{A}) = m = \text{spark}(\mathbf{A}) - 1$ with probability 1.

In general, $\forall$ full rank matrix $\mathbf{A} \in \mathbb{R}^{m \times n} (m < n)$, any $m + 1$ columns of $\mathbf{A}$ is linearly dependent, so

$$\text{spark}(\mathbf{A}) \leq m + 1 = \text{rank}(\mathbf{A}) + 1.$$

Coherence

Definition (Mallat and Zhang [1993])

The (mutual) coherence of a given matrix $\mathbf{A}$ is the largest absolute normalized inner product between different columns from $\mathbf{A}$. Suppose $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_n]$. The mutual coherence of $\mathbf{A}$ is given by

$$\mu(\mathbf{A}) = \max_{k,j,k\neq j} \frac{|\mathbf{a}_k^\top \mathbf{a}_j|}{\|\mathbf{a}_k\|_2 \cdot \|\mathbf{a}_j\|_2}.$$

- It characterizes the dependence between columns of $\mathbf{A}$
- For unitary matrices, $\mu(\mathbf{A}) = 0$
- For matrices with more columns than rows, $\mu(\mathbf{A}) > 0$
- For recovery problems, we desire a small $\mu(\mathbf{A})$ as it is similar to unitary matrices.
- For $\mathbf{A} = [\Phi\ \Psi]$ where Φ and Ψ are $n \times n$ unitary, it holds $n^{-1/2} \leq \mu(\mathbf{A}) \leq 1$
- $\mu(\mathbf{A}) = n^{-1/2}$ is achieved with $[\mathbf{I}\ \mathcal{F}]$, $[\mathbf{I}\ \mathrm{Hadamard}]$, etc.
- if $\mathbf{A} \in \mathbb{R}^{m\times n}$ where $n > m$, then $\mu(\mathbf{A}) \geq m^{-1/2}$.

Coherence

Theorem (Donoho and Elad [2003])

$$\mathrm{spark}(\mathbf{A}) \geq 1 + \mu^{-1}(\mathbf{A}).$$

Proof sketch:

- $\bar{\mathbf{A}} \leftarrow \mathbf{A}$ with columns normalized to unit 2-norm
- $p \leftarrow \mathrm{spark}(\mathbf{A})$
- $\mathbf{B} \leftarrow$ a $p \times p$ minor of $\bar{\mathbf{A}}^\top \bar{\mathbf{A}}$
- $|B_{ii}| = 1$ and $\sum_{j \neq i} |B_{ij}| \leq (p-1)\mu(\mathbf{A})$
- Suppose $p < 1 + \mu^{-1}(\mathbf{A}) \Rightarrow 1 > (p-1)\mu(\mathbf{A}) \Rightarrow |B_{ii}| > \sum_{j \neq i} |B_{ij}|, \forall i$
- $\Rightarrow \mathbf{B} \succ 0$ (Gershgorin circle theorem) $\Rightarrow \mathrm{spark}(\mathbf{A}) > p$. Contradiction.

Coherence-base guarantee

Corollary

If $\mathbf{A}\mathbf{x} = \mathbf{b}$ *has a solution* $\mathbf{x}$ *obeying* $\|\mathbf{x}\|_0 < (1 + \mu^{-1}(\mathbf{A}))/2$*, then* $\mathbf{x}$ *is the unique sparsest solution.*

Compare with the previous

Theorem

If $\mathbf{A}\mathbf{x} = \mathbf{b}$ *has a solution* $\mathbf{x}$ *obeying* $\|\mathbf{x}\|_0 < \mathrm{spark}(\mathbf{A})/2$*, then* $\mathbf{x}$ *is the sparsest solution.*

For $\mathbf{A} \in \mathbb{R}^{m \times n}$ where $m < n$, $(1 + \mu^{-1}(\mathbf{A}))$ is at most $1 + \sqrt{m}$ but spark can be $1 + m$. spark is more useful.

Assume $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution with $\|\mathbf{x}\|_0 = k < \mathrm{spark}(\mathbf{A})/2$. It will be the unique ℓ_0 minimizer. Will it be the ℓ_1 minimizer as well? Not necessarily.

However, $\|\mathbf{x}\|_0 < (1 + \mu^{-1}(\mathbf{A}))/2$ is a sufficient condition.

Coherence-based $\ell_0 = \ell_1$

Theorem (Donoho and Elad [2003], Gribonval and Nielsen [2003])

If $\mathbf{A}$ *has normalized columns and* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *has a solution* $\mathbf{x}$ *satisfying*

$$\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \mu^{-1}(\mathbf{A})\right),$$

then this $\mathbf{x}$ *is the unique minimizer with respect to both* ℓ_0 *and* ℓ_1*.*

Proof sketch:

- Previously we know $\mathbf{x}$ is the unique ℓ_0 minimizer; let $S := \mathrm{supp}(\mathbf{x})$
- Suppose $\mathbf{y}$ is the ℓ_1 minimizer but not $\mathbf{x}$; we study $\mathbf{e} := \mathbf{y} - \mathbf{x}$
- $\mathbf{e}$ must satisfy $\mathbf{A}\mathbf{e} = 0$ and $\|\mathbf{e}\|_1 \leq 2\|\mathbf{e}_S\|_1$
- $\mathbf{A}^\top\mathbf{A}\mathbf{e} = 0 \Rightarrow |e_j| \leq (1 + \mu(\mathbf{A}))^{-1}\mu(\mathbf{A})\|\mathbf{e}\|_1, \forall j$
- the last two points together contradict the assumption

Result bottom line: allow $\|\mathbf{x}\|_0$ up to $O(\sqrt{m})$ for exact recovery

The null space of $\mathbf{A}$

- **Definition**: $\|\mathbf{x}\|_p := \left(\sum_i |x_i|^p\right)^{1/p}$.
- **Lemma:** Let $0 < p \leq 1$. If $\|(\mathbf{y}-\mathbf{x})_{\bar{S}}\|_p > \|(\mathbf{y}-\mathbf{x})_S\|_p$ then $\|\mathbf{x}\|_p < \|\mathbf{y}\|_p$.
  Proof: Let $\mathbf{e} := \mathbf{y} - \mathbf{x}$.
  $\|\mathbf{y}\|_p^p = \|\mathbf{x}+\mathbf{e}\|_p^p = \|\mathbf{x}_S + \mathbf{e}_S\|_p^p + \|\mathbf{e}_{\bar{S}}\|_p^p =$
  $\|\mathbf{x}\|_p^p + (\|\mathbf{e}_{\bar{S}}\|_p^p - \|\mathbf{e}_S\|_p^p) + (\|\mathbf{x}_S + \mathbf{e}_S\|_p^p - \|\mathbf{x}_S\|_p^p + \|\mathbf{e}_S\|_p^p)$.
  Last term is nonnegative for $0 < p \leq 1$.
  So, a sufficient condition is $\|\mathbf{e}_{\bar{S}}\|_p^p > \|\mathbf{e}_S\|_p^p$. ■
- If the condition holds for $0 < p \leq 1$, it also holds for $q \in (0, p]$.
- **Definition** (null space property $\mathrm{NSP}(k,\gamma)$). Every nonzero $\mathbf{e} \in \mathcal{N}(\mathbf{A})$ satisfies $\|\mathbf{e}_S\|_1 < \gamma\|\mathbf{e}_{\bar{S}}\|_1$ for all index sets S with $|S| \leq k$.

The null space of $\mathbf{A}$

Theorem (Donoho and Huo [2001], Gribonval and Nielsen [2003])

Basis pursuit $\min\{\|\mathbf{x}\|_1 : \mathbf{Ax} = \mathbf{b}\}$ *uniquely recovers* <u>*all* k*-sparse*</u> *vectors* $\mathbf{x}^o$ *from measurements* $\mathbf{b} = \mathbf{Ax}^o$ *if and only if* $\mathbf{A}$ *satisfies* $\mathrm{NSP}(k, 1)$.

Proof.

Sufficiency. Pick any k-sparse vector $\mathbf{x}^o$. Let $S := \mathrm{supp}(\mathbf{x}^o)$ and $\bar{S} = S^c$. For any *non-zero* $\mathbf{h} \in \mathcal{N}(\mathbf{A})$, we have $\mathbf{A}(\mathbf{x}^o + \mathbf{h}) = \mathbf{Ax}^o = \mathbf{b}$ and

$$\begin{aligned} \|\mathbf{x}^0 + \mathbf{h}\|_1 &= \|\mathbf{x}_S^0 + \mathbf{h}_S\|_1 + \|\mathbf{h}_{\bar{S}}\|_1 \\ &\geq \|\mathbf{x}_S^0\|_1 - \|\mathbf{h}_S\|_1 + \|\mathbf{h}_{\bar{S}}\|_1 \qquad (1) \\ &= \|\mathbf{x}^0\|_1 + (\|\mathbf{h}_{\bar{S}}\|_1 - \|\mathbf{h}_S\|_1)\,. \end{aligned}$$

$\mathrm{NSP}(k, 1)$ of $\mathbf{A}$ guarantees $\|\mathbf{x}^0 + \mathbf{h}\|_1 > \|\mathbf{x}^0\|_1$, so $\mathbf{x}^o$ is the unique solution.
Necessity. The inequality (1) *holds with equality if* $\mathrm{sign}(\mathbf{x}_S^o) = -\mathrm{sign}(\mathbf{h}_S)$ *and* $\mathbf{h}_S$ *has a sufficiently small scale. Therefore, basis pursuit to uniquely recovers all* k*-sparse vectors* $\mathbf{x}^o$, $\mathrm{NSP}(k, 1)$ *is also necessary.* □

The null space of $\mathbf{A}$

- Another sufficient condition (Zhang [2008]) for $\|\mathbf{x}\|_1 < \|\mathbf{y}\|_1$ is

$$\|\mathbf{x}\|_0 < \frac{1}{4}\left(\frac{\|\mathbf{y}-\mathbf{x}\|_1}{\|\mathbf{y}-\mathbf{x}\|_2}\right)^2.$$

Proof:

$$\|\mathbf{e}_S\|_1 \leq \sqrt{|S|}\|\mathbf{e}_S\|_2 \leq \sqrt{|S|}\|\mathbf{e}\|_2 = \sqrt{\|x\|_0}\|\mathbf{e}\|_2.$$

Then, the above sufficient condition $\|\mathbf{y}-\mathbf{x}\|_1 > 2\|(\mathbf{y}-\mathbf{x})_S\|_1$ is given the above inequality. ■

Null space

Theorem (Zhang [2008])

Given $\mathbf{x}$ *and* $\mathbf{b} = \mathbf{A}\mathbf{x}$,

$$\min \|\mathbf{x}\|_1 \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}$$

recovers $\mathbf{x}$ *uniquely if*

$$\|\mathbf{x}\|_0 < \min \left\{ \frac{1}{4} \frac{\|\mathbf{e}\|_1^2}{\|\mathbf{e}\|_2^2} : \mathbf{e} \in \mathcal{N}(\mathbf{A}) \setminus \{0\} \right\}.$$

Comments:

- We know $1 \leq \|\mathbf{e}\|_1 / \|\mathbf{e}\|_2 \leq \sqrt{n}$ for all $\mathbf{e} \neq 0$. The ratio is small for sparse vectors but we want it large, i.e., close to $\sqrt{n}$ and away from 1.
- Fact: in most subspaces, the ratio is away from 1
- In particular, Kashin, Garvaev, and Gluskin showed that a randomly drawn $(n-m)$-dimensional subspace $\mathcal{V}$ satisfies

  $$\frac{\|\mathbf{e}\|_1}{\|\mathbf{e}\|_2} \geq \frac{c_1 \sqrt{m}}{\sqrt{1 + \log(n/m)}}, \ \mathbf{e} \in \mathcal{V}, \mathbf{e} \neq 0$$

  with probability at least $1 - \exp(-c_0(n-m))$, where $c_0, c_1 > 0$ are independent of m and n.

Null space

Theorem (Zhang [2008])

If $\mathbf{A} \in \mathbb{R}^{m \times n}$ *is sampled from i.i.d. Gaussian or is any rank-*m *matrix such that* $\mathbf{B}\mathbf{A}^\top = 0$ *and* $\mathbf{B} \in \mathbb{R}^{(n-m) \times m}$ *is i.i.d. Gaussian, then with probability at least* $1 - \exp(-c_0(n-m))$*,* ℓ_1 *minimization recovers any sparse* $\mathbf{x}$ *if*

$$\|\mathbf{x}\|_0 < \frac{c_1^2}{4} \frac{m}{1 + \log(n/m)},$$

where c_0, c_1 *are positive constants independent of* m *and* n*.*

Comments on NSP

- NSP is *no longer necessary* if "for all k-sparse vectors" is relaxed.
- NSP is *widely used in the proofs of other guarantees*.
- NSP of order $2k$ is *necessary* for stable universal recovery.
  Consider an arbitrary decoder Δ, tractable or not, that returns a vector from the input $\mathbf{b} = \mathbf{A}\mathbf{x}^o$. If one requires Δ to be stable in the sense

$$\|\mathbf{x}^o - \Delta(\mathbf{A}\mathbf{x}^o)\|_1 < C \cdot \sigma_{[k]}(\mathbf{x}^o)$$

  for all $\mathbf{x}^o$ and $\sigma_{[k]}$ is the best k-term approximation error, then it holds

$$\|\mathbf{h}_{\mathcal{S}}\|_1 < C \cdot \|\mathbf{h}_{\mathcal{S}^c}\|_1,$$

  for all non-zero $\mathbf{h} \in \mathcal{N}(\mathbf{A})$ and all coordinate sets S with $|S| \leq 2k$. See Cohen, Dahmen, and DeVore [2006].

Restricted isometry property (RIP)

Definition (Candes and Tao [2005])

Matrix $\mathbf{A}$ obeys the restricted isometry property (RIP) with constant δ_s if

$$(1-\delta_s)\|\mathbf{c}\|_2^2 \leq \|\mathbf{A}\mathbf{c}\|_2^2 \leq (1+\delta_s)\|\mathbf{c}\|_2^2$$

for all s-sparse vectors $\mathbf{c}$.

RIP essentially requires that every set of columns with cardinality less than or equal to s behaves like an orthonormal system.

RIP

Theorem (Candes and Tao [2006])

If $\mathbf{x}$ is k-sparse and $\mathbf{A}$ satisfies $\delta_{2k} + \delta_{3k} < 1$, then $\mathbf{x}$ is the unique ℓ_1 minimizer.

Comments:

- RIP needs a matrix to be properly scaled
- the tight RIP constant of a *given matrix* $\mathbf{A}$ is difficult to compute
- the result is universal for all k-sparse
- $\exists$ tighter conditions (see next slide)
- all methods (including ℓ_0) require $\delta_{2k} < 1$ for universal recovery; every k-sparse x is unique if $\delta_{2k} < 1$
- the requirement can be satisfied by certain $\mathbf{A}$ (e.g., whose entries are i.i.d samples following a subgaussian distribution) and lead to exact recovery for $\|\mathbf{x}\|_0 = O(m/\log(m/k))$.

More Comments

- (Foucart-Lai) If $\delta_{2k+2} < 1$, then $\exists$ a sufficiently small p so that ℓ_p minimization is guaranteed to recovery any k-sparse x
- (Candes) $\delta_{2k} < \sqrt{2} - 1$ is sufficient
- (Foucart-Lai) $\delta_{2k} < 2(3 - \sqrt{2})/7 \approx 0.4531$ is sufficient
- RIP gives $\kappa(\mathbf{A}_S) \leq \sqrt{(1+\delta_k)/(1-\delta_k)}$, $\forall |S| \leq k$; so $\delta_{2k} < 2(3 - \sqrt{2})/7$ gives $\kappa(\mathbf{A}_S) \leq 1.7, \ \forall |S| \leq 2m$, very well-conditioned.
- (Mo-Li) $\delta_{2k} < 0.493$ is sufficient
- (Cai-Wang-Xu) $\delta_k < 0.307$ is sufficient
- (Cai-Zhang) $\delta_k < 1/3$ is sufficient and necessary for universal ℓ_1 recovery

Random matrices with RIPs

Trivial randomly constructed matrices satisfy RIPs with overwhelming probability.

- Gaussian: $A_{ij} \sim N(0, 1/m)$, $\|\mathbf{x}\|_0 \le O(m/\log(n/m))$ whp, proof is based on applying concentration of measures to the singular values of Gaussian matrices (Szarek-91,Davidson-Szarek-01).
- Bernoulli: $A_{ij} \sim \pm 1$ wp $1/2$, $\|\mathbf{x}\|_0 \le O(m/\log(n/m))$ whp, proof is based on applying concentration of measures to the smallest singular value of a subgaussian matrix (Candes-Tao-04,Litvak-Pajor-Rudelson-TomczakJaegermann-04).
- Fourier ensemble: $A \in \mathbb{C}^{m \times n}$ is a randomly chosen submatrix of discrete Fourier transform $F \in \mathbb{C}^{n \times n}$. Candes-Tao shows $\|\mathbf{x}\|_0 \le O(m/\log(n)^6)$ whp; Rudelson-Vershynin shows $\|\mathbf{x}\|_0 \le O(m/\log(n)^4)$; conjectured $\|\mathbf{x}\|_0 \le O(m/\log(n))$.
- ......

Incoherent Sampling

Suppose (Φ, Ψ) is a pair of orthonormal bases of $\mathbb{R}^n$.

- Φ is used for sensing: $\mathbf{A}$ is a subset of rows of Φ^*
- Ψ is used to sparsely represent $\mathbf{x}$: $\mathbf{x} = \Psi\alpha$, α is sparse

Definition

The coherence between Φ and Ψ is

$$\mu(\Phi, \Psi) = \sqrt{n} \max_{1 \le k,j \le n} |\langle \phi_k, \psi_j \rangle|$$

Coherence is the largest correlation between any two elements of Φ and Ψ.

- If Φ and Ψ contains correlated elements, then $\mu(\Phi, \Psi)$ is large
- Otherwise, $\mu(\Phi, \Psi)$ is small

From linear algebra, $1 \le \mu(\Phi, \Psi) \le \sqrt{n}$.

Incoherent Sampling

Compressive sensing requires *low coherent* pairs.

- $\mathbf{x}$ is sparse under Ψ: $\mathbf{x} = \Psi\alpha$, α is sparse
- $\mathbf{x}$ is measured as $\mathbf{b} \leftarrow \mathbf{A}\mathbf{x}$
- $\mathbf{x}$ is recovered from $\min \|\alpha\|_1,$ s.t. $\mathbf{A}\Psi\alpha = \mathbf{b}$

Examples:

- Φ is spike basis $\phi_k(t) = \delta(t-k)$ and Ψ is the Fourier basis $\psi_j(t) = n^{-1/2}e^{i\cdot 2\pi\cdot jt/n}$; then $\mu(\Phi,\Psi) = 1$, achieving max incoherence.
- Coherence between noiselets and Haar wavelets is $\sqrt{2}$.
- Coherence between noiselets and Baubechies D4 and D8 are ~ 2.2 and 2.9, respectively.
- Random matrices are largely incoherent with any fixed basis Ψ. Randomly generated and orthonormalized Φ: w.h.p., the coherence between Φ and any fixed Ψ is about $\sqrt{2\log n}$.
- Similar results apply to random Gaussian or ± 1 matrices. Bottom line: many random matrices are universally incoherent with any fixed Ψ w.h.p.
- some kind of random circulant matrix is universally incoherent with any fixed Ψ w.h.p.

Incoherent Sampling

Theorem (Candes and Romberg [2007])

Fix $\mathbf{x}$ and suppose $\mathbf{x}$ is k-sparse under basis Ψ with coefficients in uniformly random signs. Select m measurements in the Φ domain uniformly at random. If $m \geq O(\mu^2(\Phi, \Psi)k \log(n))$, ℓ_1-minimization recovers $\mathbf{x}$ with high probability.

Comments:

- The result is not universal for all Ψ or all k-sparse $\mathbf{x}$ under Ψ.
- Only guaranteed for nearly all sign sequences $\mathbf{x}$ with a fixed support.
- Why seeing probability? Because there are special signals that are sparse in Ψ yet vanish at most places in the Φ domain.
- This result allows structured, as opposed to noise-like (random), matrices.
- Can be seen as an extension to Fourier CS.
- Bottom line: the smaller the coherence, the fewer the samples required. This matches numerical experience.

Robust Recovery

In order to be **practically powerful**, CS must deal with

- nearly sparse signals
- measurement noise
- sometimes both

Goal: To obtain accurate reconstructions from highly undersampled measurements, or in short, stable recovery.

Stable ℓ_1 Recovery

Consider

- a **sparse** $\mathbf{x}$
- **noisy** CS measurements $\mathbf{b} \leftarrow \mathbf{A}\mathbf{x} + \mathbf{z}$, where $\|\mathbf{z}\|_2 \leq \epsilon$

Apply the BPDN model: $\min \|\mathbf{x}\|_1$ s.t. $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \epsilon$.

Theorem

Assume (some bounds on δ_k or δ_{2k}). The solution of the BPDN model returns a solution $\mathbf{x}^*$ satisfying

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \leq C \cdot \epsilon$$

for some constant C.

Proof sketch (using an overly sufficient RIP bound):

- Let $\mathbf{e} = \mathbf{x}^* - \mathbf{x}$. $S' = \{i : \text{largest } 2k\ |x|_{(i)}\}$.
- One can show $\|\mathbf{e}\|_2 \leq C_1 \|\mathbf{e}_{S'}\|_2 \leq C_2 \|\mathbf{A}\mathbf{e}\|_2 \leq C \cdot \epsilon$.
  - 1st inequality essentially from $\|\mathbf{e}_S\|_1 > \|\mathbf{e}_{\bar{S}}\|_1$,
  - 2nd inequality essentially from the RIP;
  - 3rd inequality essentially from the constraint.

Stable ℓ_1 Recovery

Theorem

Assume (some bounds on δ_k or δ_{2k}). The solution of the BPDN model returns a solution $\mathbf{x}^*$ satisfying

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \leq C \cdot \epsilon$$

for some constant C.

Comments:

- The result is universal and more general than exact recovery;
- The error bound is order-optimal: knowing $\mathrm{supp}(\mathbf{x})$ will give $C' \cdot \epsilon$ at best;
- $\mathbf{x}^*$ is almost as good as if one knows where the largest k entries are and directly measure them;
- C depends on k; when k violates the condition and gets too large, $\|\mathbf{x}^* - \mathbf{x}\|_2$ will blow up.

Stable ℓ_1 Recovery

Consider

- a nearly sparse $\mathbf{x} = \mathbf{x}_k + \mathbf{w}$,
- $\mathbf{x}_k$ is the vector $\mathbf{x}$ with all but the largest (in magnitude) k entries set to 0,
- CS measurements $\mathbf{b} \leftarrow \mathbf{A}\mathbf{x} + \mathbf{z}$, where $\|\mathbf{z}\|_2 \leq \epsilon$.

Theorem

Assume (some bounds on δ_k or δ_{2k}). The solution of the BPDN model returns a solution $\mathbf{x}^*$ satisfying

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \leq \bar{C} \cdot \epsilon + \tilde{C} \cdot k^{-1/2} \|\mathbf{x} - \mathbf{x}_k\|_1$$

for some constants $\bar{C}$ and $\tilde{C}$.

Proof sketch (using an overly sufficient RIP bound): Similar to the previous one, **except** x is no longer k-sparse and $\|\mathbf{e}_S\|_1 > \|\mathbf{e}_{\bar{S}}\|_1$ is no longer valid. Instead, we get $\|\mathbf{e}_S\|_1 + 2\|\mathbf{x} - \mathbf{x}_k\|_1 > \|\mathbf{e}_{\bar{S}}\|_1$. Then,
$\|\mathbf{e}\|_2 \leq C_1 \|\mathbf{e}_{4k}\|_2 + C' k^{-1/2} \|\mathbf{x} - \mathbf{x}_k\|_1$, ...

Stable ℓ_1 Recovery

Comments on

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \le \bar{C} \cdot \epsilon + \tilde{C} \cdot k^{-1/2} \|\mathbf{x} - \mathbf{x}_k\|_1.$$

Suppose $\epsilon = 0$; let us focus on the last term:

1. Consider power-law decay signals $|x|_{(i)} \le C \cdot i^{-r}$, $r > 1$;
2. Then, $\|\mathbf{x} - \mathbf{x}_k\|_1 \le C_1 k^{-r+1}$ or $k^{-1/2}\|\mathbf{x} - \mathbf{x}_k\|_1 \le C_1 k^{-r+(1/2)}$;
3. But even if $\mathbf{x}^* = \mathbf{x}_k$, $\|\mathbf{x}^* - \mathbf{x}\|_2 = \|\mathbf{x}_k - \mathbf{x}\|_2 \le C_1 k^{-r+(1/2)}$;
4. Conclusion: the bound cannot be fundamentally improved.

Information Theoretic Analysis

Question: is there an encoding-decoding means that can do *fundamentally better* than Gaussian $\mathbf{A}$ and ℓ_1-minimization?

In math: $\exists$ encoder-decoder pair $(\mathbf{A}, \Delta)$, $\ni \|\Delta(\mathbf{A}\mathbf{x}) - \mathbf{x}\|_2 \leq O(k^{-1/2}\sigma_k(\mathbf{x}))$ holds for k *larger than* $O(m/\log(n/m))$?

Comments: $\mathbf{A}$ can be any matrix, and Δ can be *any decoder*, tractable or not.

Let $\|\mathbf{x} - \mathbf{x}_k\|_1$ be called the best-k approximation error, denoted by $\sigma_k(\mathbf{x}) := \|\mathbf{x} - \mathbf{x}_k\|_1$.

Performance of $(\mathbf{A}, \Delta)$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$:

$$E_m(K) := \inf_{(\mathbf{A},\Delta)} \sup_{\mathbf{x} \in K} \|\Delta(\mathbf{A}\mathbf{x}) - \mathbf{x}\|_2$$

Gelfand width:

$$d^m(K) = \inf_{\operatorname{codim}(Y) \le m} \sup \left\{ \|\mathbf{h}\|_2 : \mathbf{h} \in K \cap Y \right\}.$$

Cohen, Dahmen, and DeVore [2006]: If set $K = -K$ and $K + K \le C_0 K$, then

$$d^m(K) \le E_m(K) \le C_0 d^m(K).$$

Gelfand Width and $K = \ell_1$-ball

Kashin, Gluskin, Garnaev: for $K = \{\mathbf{h} \in \mathbb{R}^n : \|\mathbf{h}\|_1 \leq 1\}$,

$$C_1\sqrt{\frac{\log(n/m)}{m}} \leq d^m(K) \leq C_2\sqrt{\frac{\log(n/m)}{m}}.$$

Consequences:

1. KGG means $E_m(K) \approx \sqrt{\frac{\log(n/m)}{m}}$
2. we want $\|\Delta(\mathbf{Ax}) - \mathbf{x}\|_2 \leq C \cdot k^{-1/2}\sigma_k(x) \leq C \cdot k^{-1/2}\|x\|_1$; normalizing gives $E_m(K) \leq C \cdot k^{-1/2}$.
3. Therefore, $k \leq m/\log(n/m)$. We cannot do better than this.

(Incomplete) References:

D. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries vis ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100:2197–2202, 2003.

R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

E. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

Y. Zhang. Theory of compressive sensing via l1-minimization: a non-RIP analysis and extensions. *Rice University CAAM Technical Report TR08-11*, 2008.

I.F. Gorodnitsky and B.D. Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *Signal Processing, IEEE Transactions on*, 45(3):600–616, 1997.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

D. Donoho and X. Huo. Uncertainty principles and ideal atomic decompositions. *IEEE Transactions on Information Theory*, 47:2845–2862, 2001.

A. Cohen, W. Dahmen, and R. A. DeVore. Compressed sensing and best k-term approximation. *Submitted.*, 2006.

E. Candes and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies. *IEEE Transactions on Information Theory*, 52(1): 5406–5425, 2006.

E. Candes and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.