Lecture: Introduction to Compressed Sensing
Sparse Recovery Guarantees

http://bicmr.pku.edu.cn/~wenzw/bigdata2015.html

Acknowledgement: this slides is based on Prof. Emmanuel Candes’ and Prof. Wotao Yin’s lecture notes
Underdetermined systems of linear equations

- \( x \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m \)

When fewer equations than unknowns

- Fundamental theorem of algebra says that we cannot find \( x \)
- In general, this is absolutely correct
Special structure

If unknown is assumed to be
- sparse
- low-rank

then one can *often* find solutions to these problems by convex optimization
Compressive Sensing

Find the sparsest solution

- Given $n=256$, $m=128$.
- $A = \text{randn}(m, n)$; $u = \text{sprandn}(n, 1, 0.1)$; $b = A^*u$;

\[
\begin{align*}
\min_x \|x\|_0 & \quad \text{s.t. } Ax = b \\
\min_x \|x\|_2 & \quad \text{s.t. } Ax = b \\
\min_x \|x\|_1 & \quad \text{s.t. } Ax = b
\end{align*}
\]

(a) $\ell_0$-minimization  
(b) $\ell_2$-minimization  
(c) $\ell_1$-minimization
Linear programming formulation

\[ \min \|x\|_{\ell_0} \quad \text{s.t.} \quad Ax = b \]

Combinatorially hard

\[ \min \|x\|_{\ell_1} \quad \text{s.t.} \quad Ax = b \]

Linear program

\[ \begin{align*}
\text{minimize} & \quad \sum_i |x_i| \\
\text{subject to} & \quad Ax = b
\end{align*} \]

is equivalent to

\[ \begin{align*}
\text{minimize} & \quad \sum_i t_i \\
\text{subject to} & \quad Ax = b \\
& \quad -t_i \leq x_i \leq t_i
\end{align*} \]

with variables \( x, t \in \mathbb{R}^n \)

\( x^* \text{ solution} \iff (x^*, t^* = |x^*|) \text{ solution} \)
Compressed sensing

- Name coined by David Donoho
- Has become a label for sparse signal recovery
- But really one instance of underdetermined problems

- Informs analysis of underdetermined problems
- Changes viewpoint about underdetermined problems
- Starting point of a general burst of activity in
  - information theory
  - signal processing
  - statistics
  - some areas of computer science
  - ...

- Inspired new areas of research, e.g. low-rank matrix recovery
A contemporary paradox

- Massive data acquisition
- Most of the data is redundant and can be thrown away
- Seems enormously wasteful
Sparsity in signal processing

**Implication:** can discard small coefficients without much perceptual loss
Sparsity and wavelet "compression"

Take a mega-pixel image

- Compute 1,000,000 wavelet coefficients
- Set to zero all but the 25,000 largest coefficients
- Invert the wavelet transform

This principle underlies modern lossy coders
Comparison

Sparse representation = good compression
Why? Because there are fewer things to send/store

Traditional
Compressive sensing

Compressive sensing (senses less, faster)
How many measurements to acquire a sparse signal?

- $x$ is $s$-sparse
- Take $m$ random and nonadaptive measurements, e.g. $a_k \sim \mathcal{N}(0, I)$:
  \[
y_k = \langle a_k, x \rangle, \quad k = 1, \ldots, m
  \]
- Reconstruct by $\ell_1$ minimization

First fundamental result

if $m \geq s \log n$

- Recovers original exactly
- Efficient acquisition is possible by nonadaptive sensing

Cannot do essentially better with

- fewer measurements
- with other reconstruction algorithms
• signal is local
• measurements are global
Signals/images may not be exactly sparse

image

wavelet coefficient table
Nonadaptive sensing of compressible signals

Classical viewpoint
- Measure everything (all the pixels, all the coefficients)
- Store $s$-largest coefficients

$$\text{distortion} = \|x - x_s\|_{\ell_2}$$

Compressed sensing
- Take $m$ random measurements
$$b_k = \langle a_k, x \rangle$$
- Reconstruct by $\ell_1$ minimization

Second fundamental result
With $m \geq s \log n$ or even $m \geq s \log n / s$

$$\|\hat{x} - x\|_{\ell_2} \leq \|x - x_s\|_{\ell_2}$$

Simultaneous (nonadaptive) acquisition and compression
But data are always noisy...

Random sensing with $a \sim \mathcal{N}(0, I)$

$$y_k = \langle a_k, x \rangle + \delta z_k, \quad k = 1, \ldots, m$$

- $a \sim \mathcal{N}(0, I)$ (say)
- $z_k \text{ iid } \mathcal{N}(0, 1)$ (say)

Third fundamental result: C. and Plan ('09)

- Recovery via lasso or Dantzig selector
- $\bar{s} = m / \log(n/m)$

$$||\hat{x} - x||^2_{\ell_2} \leq \inf_{1 \leq s \leq \bar{s}} ||x - x_s||^2_{\ell_2} + \log n \frac{s \delta^2}{m}$$

= near-optimal bias-variance trade off

Result holds more generally
Oppportunities

When measurements are

- expensive (e.g. fuel cell imaging, near IR imaging)
- slow (e.g. MRI)
- beyond current capabilities (e.g. wideband analog to digital conversion)
- wasteful
- missing
- ...

Compressed sensing may offer a way out
Fundamental Question

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?
How to read guarantees

Some basic aspects that distinguish different types of guarantees:

- Recoverability (exact) vs stability (inexact)
- General A or special A?
- Universal (all sparse vectors) or instance (certain sparse vector(s))?
- General optimality? or specific to model/algorithm?
- Required property of 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?
First questions for finding the sparsest solution to \( Ax = b \)

- Can sparsest solution be unique? Under what conditions?
- Given a sparse \( x \), how to verify whether it is actually the sparsest one?

**Definition (Donoho and Elad 2003)**

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

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

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

Theorem (Gorodnitsky and Rao 1997)

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

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

\[
\|x\|_0 + \|y\|_0 \geq \|x - y\|_0 \geq \text{spark}(A),
\]

or \(\|y\|_0 \geq \text{spark}(A) - \|x\|_0 > \text{spark}(A)/2 > \|x\|_0\)

The result does not mean this \(x\) can be efficiently found numerically.

For many random matrices \(A \in \mathbb{R}^{m \times n}\), the result means that if an algorithm returns \(x\) satisfying \(\|x\|_0 < (m + 1)/2\), the \(x\) is optimal with probability 1.

What to do when \(\text{spark}(A)\) is difficult to obtain?
Rank is easy to compute, but spark needs a combinatorial search.

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

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

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

\[
\text{spark}(A) \leq m + 1 = \text{rank}(A) + 1
\]
Coherence

**Definition (Mallat and Zhang 1993)**

The (mutual) coherence of a given matrix $A$ is the largest absolute normalized inner product between different columns from $A$. Suppose $A = [a_1, a_2, \ldots, a_n]$. The mutual coherence of $A$ is given by

$$
\mu(A) = \max_{k,j, k \neq j} \frac{|a_k^\top a_j|}{\|a_k\|_2 \cdot \|a_j\|_2}
$$

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

Theorem (Donoho and Elad 2003)

$$\text{spark}(A) \geq 1 + \mu^{-1}(A)$$

Proof Sketch

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

**Corollary**

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

Compare with the previous

**Theorem (Gorodnitsky and Rao 1997)**

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

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

- Assume \( Ax = b \) has a solution with \( \| x \|_0 = k < \text{spark}(A)/2 \). It will be the unique \( \ell_0 \) minimizer. Will it be the \( \ell_1 \) minimizer as well? Not necessarily. However, \( \| x \|_0 < (1 + \mu^{-1}(A))/2 \) is a sufficient condition.
Coherence-based $\ell_0 = \ell_1$

**Theorem (Donoho and Elad 2003, Gribonval and Nielsen 2003)**

If $A$ has normalized columns and $Ax = b$ has a solution $x$ satisfying $\|x\|_0 \leq (1 + \mu^{-1}(A))/2$, then $x$ is the unique minimizer with respect to both $\ell_0$ and $\ell_1$.

**Proof Sketch**

- Previously we know $x$ is the unique $\ell_0$ minimizer; let $S := \text{supp}(x)$
- Suppose $y$ is the $\ell_1$ minimizer but not $x$; we study $e := y - x$
- $e$ must satisfy $Ae = 0$ and $\|e\|_1 \leq 2\|e_S\|_1$ since $0 \geq \|y\|_1 - \|x\|_1 = \sum_{i \in S^c} |y_i| - \sum_{i \in S} (|y_i| - |x_i|) \geq \|e_{S^c}\|_1 - \sum_{i \in S} |y_i - x_i| = \|e_{S^c}\|_1 - \|e_S\|_1$
- $A^\top A e = 0 \Rightarrow |e_j| \leq (1 + \mu(A))^{-1} \mu(A) \|e\|_1$, $\forall j$
- the last two points together contradict the assumption

Result bottom line: allow $\|x\|_0$ up to $O(\sqrt{m})$ for exact recovery
The null space of $A$

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

- **Lemma**: Let $0 < p < 1$. If $\|(y - x)_{Sc}\|_p > \|(y - x)_S\|_p$, then $\|x\|_p < \|y\|_p$.
  **Proof**: Let $e = y - x$.
  \[
  \|y\|_p^p = \|x + e\|_p^p = \|x_S + e_S\|_p^p + \|e_{Sc}\|_p^p = \|x\|_p^p + (\|e_{Sc}\|_p^p - \|e_S\|_p^p) + (\|x_S + e_S\|_p^p - \|x_S\|_p^p + \|e_S\|_p^p))
  \]
  The last term is nonnegative for $0 < p < 1$. Hence, a sufficient condition is $\|e_{Sc}\|_p^p > \|e_S\|_p^p$.

- If the condition holds for $0 < p < 1$, it also holds for $q \in (0, p]$.

- **Definition** (null space property $NSP(k, \gamma)$). Every nonzero $e \in \mathcal{N}(A)$ satisfies $\|e_S\|_1 < \gamma \|e_{Sc}\|_1$ for all index sets $S$ with $|S| \leq k$. 

The null space of $A$

**Theorem (Donoho and Huo 2001, Gribonval and Nielsen 2003)**

$$\min \|x\|_1, \ s.t. \ Ax = b \text{ uniquely recovers all } k\text{-sparse vectors } x^o \text{ from measurements } b = Ax^o \text{ if and only if } A \text{ satisfies } NSP(k, 1).$$

**Proof:**

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

$$\|x^0 + h\|_1 = \|x^o_S + h_S\|_1 + \|h_{\bar{S}}\|_1$$

$$\geq \|x^o_S\|_1 - \|h_S\|_1 + \|h_{\bar{S}}\|_1$$

$$= \|x^o_S\|_1 - (\|h_S\|_1 - \|h_{\bar{S}}\|_1)$$

- **Necessity.** The inequality holds with equality if $\text{sgn}(x^o_S) = -\text{sgn}(h_S)$ and $h_S$ has a sufficiently small scale. Therefore, basis pursuit to uniquely recovers all $k$-sparse vectors $x^o$, $NSP(k, 1)$ is also necessary.
The null space of A

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

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

- Proof:

  \[
  \|e_S\|_1 \leq \sqrt{|S|} \|e_S\|_2 \leq \sqrt{|S|} \|e\|_2 = \sqrt{\|x\|_0} \|e\|_2.
  \]

  Then, the above inequality and the sufficient condition gives

  \[
  \|y - x\|_1 > 2 \| (y - x)_S \|_1 \text{ which is } \| (y - x)_{S^c} \|_1 > \| (y - x)_S \|_1.
  \]

**Theorem (Zhang, 2008)**

\[
\min \|x\|_1, \ s.t. \ Ax = b \text{ recovers } x \text{ uniquely if}
\]

\[
\|x\|_0 < \min \left\{ \frac{1}{4} \left( \frac{\|e\|_1}{\|e\|_2} \right)^2, \quad e \in \mathcal{N}(A) \setminus \{0\} \right\}
\]
1 \leq \frac{\|v\|_1}{\|v\|_2} \leq \sqrt{n}, \quad \forall v \in \mathbb{R}^n \setminus \{0\}

Garnaev and Gluskin established that for any natural number \( p < n \), there exist \( p \)-dimensional subspaces \( V_p \subset \mathbb{R}^n \) in which

\[
\frac{\|v\|_1}{\|v\|_2} \geq \frac{C \sqrt{n - p}}{\sqrt{\log(n/(n - p))}}, \quad \forall v \in V_p \setminus \{0\},
\]

vectors in the null space of \( A \) will satisfy, with high probability, the Garnaev and Gluskin inequality for \( V_p = \text{Null}(A) \) and \( p = n - m \).

for a random Gaussian matrix \( A \), \( \bar{x} \) will uniquely solve \( \ell_1 \)-min with high probability whenever

\[
\|\bar{x}\|_0 < \frac{C^2}{4} \frac{m}{\log(n/m)}.
\]
Definition (Restricted isometry constants)

For each \( k = 1, 2, \ldots \), \( \delta_k \) is the smallest scalar such that

\[
(1 - \delta_k) \|x\|_{\ell_2}^2 \leq \|Ax\|_{\ell_2}^2 \leq (1 + \delta_k) \|x\|_{\ell_2}^2
\]

for all \( k \)-sparse \( x \)

- Note slight change of normalization
- When \( \delta_k \) is not too large, condition says that all \( m \times k \) submatrices are well conditioned (sparse subsets of columns are not too far from orthonormal)
Interlude: when does sparse recovery make sense?

- $x$ is $s$-sparse: $\|x\|_{\ell_0} \leq s$
- can we recover $x$ from $Ax = b$?

Perhaps possible if sparse vectors lie away from null space of $A$

Yes if any $2s$ columns of $A$ are linearly independent

If $x_1, x_2$ $s$-sparse such that $Ax_1 = Ax_2 = b$

$A(x_1 - x_2) = 0 \Rightarrow x_1 - x_2 = 0 \iff x_1 = x_2$
Interlude: when does sparse recovery make sense?

- $x$ is $s$-sparse: $\|x\|_{\ell_0} \leq s$
- can we recover $x$ from $Ax = b$?

Perhaps possible if sparse vectors lie away from null space of $A$

In general, **No** if $A$ has $2s$ linearly dependent columns

$h \neq 0$ is $2s$-sparse with $Ah = 0$

$h = x_1 - x_2$  $x_1, x_2$ both $s$-sparse

$Ah = 0 \iff Ax_1 = Ax_2$ and $x_1 \neq x_2$
Equivalent view of restricted isometry property

\( \delta_{2k} \) is the smallest scalar such that

\[
(1 - \delta_{2k}) \|x_1 - x_2\|_2^2 \leq \|Ax_1 - Ax_2\|_2^2 \leq (1 + \delta_{2k}) \|x_1 - x_2\|_2^2
\]

for all \( k \)-sparse vectors \( x_1, x_2 \).

The positive lower bounds is that which really matters

- If lower bound does not hold, then we may have \( x_1 \) and \( x_2 \) both sparse and with disjoint supports, obeying

\[
Ax_1 = Ax_2
\]

- Lower bound guarantees that distinct sparse signals cannot be mapped too closely (analogy with codes)
With a picture

For all $k$-sparse $x_1$ and $x_2$

\[ 1 - \delta_{2k} \leq \frac{{\|Ax_1 - Ax_2\|}^2}{\|x_1 - x_2\|^2} \leq 1 + \delta_{2k} \]
Formal equivalence

Suppose there is an $s$-sparse solution to $Ax = b$

- $\delta_{2s} < 1$ solution to combinatorial optimization ($\min \ell_0$) is unique
- $\delta_{2s} < 0.414$ solution to LP relaxation is unique and the same

Comments:

- RIP needs a matrix to be properly scaled
- the tight RIP constant of a given matrix $A$ is difficult to compute
- the result is universal for all $s$-sparse
- $\exists$ tighter conditions (see next slide)
- all methods (including $\ell_0$) require $\delta_{2s} < 1$ for universal recovery; every $s$-sparse $x$ is unique if $\delta_{2s} < 1$
- the requirement can be satisfied by certain $A$ (e.g., whose entries are i.i.d samples following a subgaussian distribution) and lead to exact recovery for $\|x\|_0 = O(m/\log(m/k))$. 
(Foucart-Lai) If $\delta_{2s+2} < 1$, then $\exists$ a sufficiently small $p$ so that $\ell_p$ minimization is guaranteed to recovery any $s$-sparse $x$

(Candes) $\delta_{2s} < \sqrt{2} - 1$ is sufficient

(Foucart-Lai) $\delta_{2s} < 2(3 - \sqrt{2})/7 \sim 0.4531$ is sufficient

RIP gives $\kappa(A_S) \leq \sqrt{(1 + \delta_s)/(1 - \delta_s)}$, $\forall |S| \leq k$. so $\delta_{2s} < 2(3 - \sqrt{2})/7$ gives $\kappa(A_S) \leq 1.7$, $\forall |S| \leq 2m$, very well-conditioned.

(Mo-Li) $\delta_{2s} < 0.493$ is sufficient

(Cai-Wang-Xu) $\delta_{2s} < 0.307$ is sufficient

(Cai-Zhang) $\delta_{2s} < 1/3$ is sufficient and necessary for universal $\ell_1$ recovery
Characterization of $\ell_1$ solutions

Underdetermined system: $A \in \mathbb{R}^{m \times n}, m < n$

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \quad \text{s.t.} \quad Ax = b$$

$x$ is solution iff

$$\|x + h\|_1 \geq \|x\|_1 \quad \forall h \in \mathbb{R}^n \quad \text{s.t.} \quad Ah = 0$$

Notations: $x$ supported on $T = \{i : x_i \neq 0\}$

$$\|x + h\|_1 = \sum_{i \in T} |x_i + h_i| + \sum_{i \in T^c} |h_i|$$

$$\geq \sum_{i \in T} |x_i| + \sum_{i \in T} \text{sgn}(x_i)h_i + \sum_{i \in T^c} |h_i|$$

because $|x_i + h_i| \geq |x_i| + \text{sgn}(x_i)h_i$

Necessary and sufficient condition for $\ell_1$ recovery

For all $h \in \text{null}(A)$

$$\sum_{i \in T} \text{sgn}(x_i)h_i \leq \sum_{i \in T^c} |h_i|$$
Why is this necessary? If there is \( h \in \text{null}(A) \) with

\[
\sum_{i \in T} \text{sgn}(x_i)h_i > \sum_{i \in T^c} |h_i|
\]

then

\[
\|x - h\|_{\ell_1} < \|x\|_{\ell_1}.
\]

**Proof:** there exists a small enough \( t \) such that

\[
|x_i + th_i| = \begin{cases} 
  x_i + th_i = x_i + t\text{sgn}(x_i)h_i & \text{if } x_i > 0 \\
  -(x_i + th_i) = -x_i + t\text{sgn}(x_i)h_i & \text{if } x_i < 0 \\
  t|h_i| & \text{otherwise}
\end{cases}
\]
Geometric picture

Null Space condition

\[ \sum_{i \in T} \text{sgn}(x_i)h_i \leq \sum_{i \in T^c} |h_i|, \ \forall h \in \text{null}(A) \]

if in addition, 

\[ \sum_{i \in T} \text{sgn}(x_i)h_i < \sum_{i \in T^c} |h_i|, \ \forall h \neq 0 \in \text{null}(A) \]

then \( x \) is the unique solution

Recovery property only depends on the sign pattern of \( x \), not the magnitudes!
Characterization via KKT conditions

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b$$

- $f$ convex and differentiable Lagrangian
- $L(x, \lambda) = f(x) + \langle \lambda, b - Ax \rangle$

$Ax = 0$ if and only if $x$ is orthogonal to each of the row vectors of $A$.

**KKT condition**

$x$ is solution iff $x$ is feasible and $\exists \lambda \in \mathbb{R}^m$ s.t.

$$\nabla_x L(x, \lambda) = 0 = \nabla f(x) - A^* \lambda$$

Geometric interpretation: $\nabla f(x) \perp \text{null}(A)$

When $f$ is not differentiable, condition becomes: $x$ feasible and $\exists \lambda \in \mathbb{R}^m$ s.t.

$A^* \lambda$ is a subgradient of $f$ at $x$
Subgradient

Definition

\( u \) is a subgradient of convex \( f \) at \( x_0 \) if for all \( x \)

\[
f(x) \geq f(x_0) + u \cdot (x - x_0)
\]

if \( f \) is differentiable at \( x_0 \), the only subgradient is \( \nabla f(x_0) \)

Subgradients of \( f(t) = |t|, t \in \mathbb{R} \)

\[
\begin{cases}
\{ \text{subgradients} \} = \{ \text{sgn}(t) \} & t \neq 0 \\
\{ \text{subgradients} \} = [-1, 1] & t = 0
\end{cases}
\]

Subgradients of \( f(x) = \|x\|_{\ell_1}, x \in \mathbb{R}^n \):

\( u \in \partial \|x\|_{\ell_1} \) (\( u \) is a subgradient) iff

\[
\begin{cases}
u_i = \text{sgn}(x_i) & x_i \neq 0 \\
|u_i| \neq 1 & x_i = 0
\end{cases}
\]
Optimality conditions II

\[
\min_{x \in \mathbb{R}^n} \|x\|_{\ell_1} \quad \text{s.t.} \quad Ax = b
\]

The dual problem is

\[
\max_y y^\top b, \quad \text{s.t.} \quad \|A^*y\|_{\infty} \leq 1
\]

\(x\) optimal solution iff there exists \(u = A^*\lambda(u \perp \text{null}(A))\) with

\[
\begin{cases}
u_i = \text{sgn}(x_i) & x_i \neq 0 \ (i \in T) \\|u_i| \leq 1 & x_i = 0 \ (i \in T^c)\end{cases}
\]

If in addition

- \(|u_i| < 1\) when \(x_i = 0\)
- \(A_T\) has full col. rank

Then \(x\) is the unique solution

Will call such a \(u\) or \(\lambda\) a dual certificate
Unicity

Notation

- $x_T$: restriction of $x$ to indices in $T$
- $A_T$: submatrix with column indices in $T$

If $\text{supp}(x) \subset T$,

$$Ax = A_T x_T.$$ 

Let $h \in \text{null}(A)$. Since $u \perp \text{null}(A)$, we have

$$\sum_{i \in T} \text{sgn}(x_i)h_i = \sum_{i \in T} u_i h_i = \langle u, h \rangle - \sum_{i \in T^c} u_i h_i$$

$$= -\sum_{i \in T^c} u_i h_i < \sum_{i \in T^c} |h_i|$$

unless $h_{T^c} \neq 0$. Now if $h_{T^c} = 0$, then since $A_T$ has full column rank,

$$Ah = A_T h_T = 0 \Rightarrow h_T = 0 \Rightarrow h = 0$$

In conclusion, for any $h \in \text{null}(A)$, $\|x + h\|_1 \geq \|x\|_1$, unless $h \neq 0$. 


Sufficient conditions

- \( T = \text{supp}(x) \) and \( A_T \) has full col rank (\( A_T^*A_T \) invertible)
- \( \text{sgn}(x_T) \) is the sign sequence of \( x \) on \( T \) and set
  \[
u := A^*A_T(A_T^*A_T)^{-1}\text{sgn}(x_T)\]

- if \( |u_i| \leq 1 \) for all \( i \in T^c \), then \( x \) is solution
- if \( |u_i| < 1 \) for all \( i \in T^c \), then \( x \) is the unique solution

Why?

- \( u \) is of the form \( A^*\lambda \)
- \( u_i = \text{sgn}(x_i) \) if \( i \in T \), since
  \[
u_T = A_T^*A_T(A_T^*A_T)^{-1}\text{sgn}(x_T) = \text{sgn}(x_T)\]

So \( u \) is a valid dual certificate
Why this special dual certificate

Solution to

\[ \begin{align*}
\text{minimize} & \quad \| u \|_2 \\
\text{subject to} & \quad u = A^* \lambda \\
& \quad u_i = \text{sgn}(x_i), i \in T
\end{align*} \]

- Explicit expression
- By minimizing 2-norm, hope to make \(|u_i|, i \in T^c\), small componentwise
- Why \(|u_i| < 1\) for all \(i \in T^c\)? Read “E. Candes and T. Tao. Decoding by linear programming. IEEE Transactions on Information Theory, 51:4203–4215, 2005”.
General setup

- $x$ not necessarily sparse
- observe $b = Ax$
- recover by $\ell_1$ minimization

$$\min \|\hat{x}\|_{\ell_1} \text{ s. t. } A\hat{x} = b$$

Interested in comparing performance with sparsest approximation $x_s$:

$$x_s = \arg \min \min_{\|z\|_{\ell_0} \leq s} \|x - z\|$$

- $x_s$: $s$-sparse
- $s$-largest entries of $x$ are the nonzero entries of $x_s$
General signal recovery

Theorem (Noiseless recovery (C., Romberg and Tao$^a$))

If $\delta_{2s} < \sqrt{2} - 1 = 0.414\ldots$, $\ell_1$ recovery obeys

$$||\hat{x} - x||_2 \lesssim ||x - x_s||_1 / \sqrt{s}$$
$$||\hat{x} - x||_1 \lesssim ||x - x_s||_1$$

- Deterministic (nothing is random)
- Universal (applies to all $x$)
- Exact if $x$ is $s$-sparse
- Otherwise, essentially reconstructs the $s$ largest entries of $x$
- Powerful if $s$ is close to $m$
General signal recovery from noisy data

Inaccurate measurements: \( z \) error term (stochastic or deterministic)

\[
b = Ax + z, \quad \text{with} \quad \|z\|_2 \leq \epsilon
\]

Recovery via the LASSO: \( \ell_1 \) minimization with relaxed constraints

\[
\min \|\hat{x}\|_1 \quad \text{s. t.} \quad \|A\hat{x} - b\|_2 \leq \epsilon
\]

**Theorem (C., Romberg and Tao)**

Assume \( \delta_{2s} < \sqrt{2} - 1 \), then

\[
\|\hat{x} - x\|_2 \lesssim \frac{\|x - x_s\|_1}{\sqrt{s}} + \epsilon = \text{approx.error} + \text{measurement error}
\]

(inequality constants hidden in \( \lesssim \) are explicit, see \( C_0 \) and \( C_1 \) on P56)

- When \( \epsilon = 0 \) (no noise), earlier result
- Says when we can solve underdetermined systems of equations accurately
Preliminaries: Lemma 1

1. If \( u \in \Sigma_k \), then \( \|u\|_1 / \sqrt{k} \leq \|u\|_2 \leq \sqrt{k}\|u\|_\infty \).
   \textbf{Proof:} \( \|u\|_1 = |\langle u, \text{sgn}(u) \rangle| \leq \|u\|_2 \|\text{sgn}(u)\|_2 \).

2. Let \( u, v \) be orthogonal vectors. Then \( \|u\|_2 + \|v\|_2 \leq \sqrt{2}\|u + v\|_2 \).
   \textbf{Proof:} Apply the first statement with \( w = (\|u\|_2, \|v\|_2)^T \).

3. Let \( A \) satisfies RIP of order \( 2k \). then for any \( x, x' \in \Sigma_k \) with disjoint supports
   \[ |\langle Ax, Ax' \rangle| \leq \delta_{s+s'} \|x\|_2 \|x'\|_2 \]
   \textbf{Proof:} Suppose \( x \) and \( x' \) are unit vectors as above. Then \( \|x + x'\|_2 = 2, \|x - x'\|_2 = 2 \) due to the disjoint supports. The RIP gives
   \[ 2(1 - \delta_{s+s'}) \leq \|Ax \pm Ax'\|_2^2 \leq 2(1 + \delta_{s+s'}) \]
   \textbf{Parallelogram identity}
   \[ |\langle Ax, Ax' \rangle| = \frac{1}{4} \|Ax + Ax'\|_2^2 - \|Ax - Ax'\|_2^2 \leq \delta_{s+s'} \]
Preliminaries: Lemma 2

Let $T_0$ be any subset $\{1, 2, \ldots, n\}$ such that $|T_0| \leq s$. For any $u \in \mathbb{R}^n$, define $T_1$ as the index set corresponding to the $s$ entries of $u_{T_0^c}$ with largest magnitude, $T_2$ as indices of the next $s$ largest coefficients, and so on. Then

$$
\sum_{j \geq 2} \|u_{T_j}\|_2 \leq \frac{\|u_{T_0^c}\|_1}{\sqrt{s}}
$$

**Proof:** We begin by observing that for $j \geq 2$,

$$
\|u_{T_j}\|_{\infty} \leq \frac{\|u_{T_{j-1}}\|_1}{s}
$$

since the $T_j$ sort $u$ to have decreasing magnitude. Using Lemma 1.1, we have

$$
\sum_{j \geq 2} \|u_{T_j}\|_2 \leq \sqrt{s} \sum_{j \geq 2} \|T_j\|_\infty \leq \sum_{j \geq 1} \frac{\|u_{T_j}\|_1}{\sqrt{s}} = \frac{\|u_{T_0^c}\|_1}{\sqrt{s}}
$$
Preliminaries: Lemma 3

Let $A$ satisfies the RIP with order $2s$. Let $T_0$ be any subset \{1, 2, \ldots, n\} such that $|T_0| \leq s$ and $h \in \mathbb{R}^n$ be given. Define $T_1$ as the index set corresponding to the $s$ entries of $h_{T_0^c}$ with largest magnitude, and set $T = T_0 \cup T_1$. Then

$$
\|h_T\|_2 \leq \alpha \frac{\|h_{T_0^c}\|_1}{\sqrt{s}} + \beta \frac{|\langle Ah_T, Ah \rangle|}{\|h_T\|_2}
$$

where $\alpha = \frac{\sqrt{2}\delta_{2s}}{1-\delta_{2s}}$ and $\beta = \frac{1}{1-\delta_{2s}}$.

Proof: Since $h_T \in \Sigma_{2s}$, the RIP gives

$$
(1 - \delta_{2s})\|h_T\|_2^2 \leq \|Ah_T\|_2^2.
$$
Continue: Proof Lemma 3

Define $T_j$ as Lemma 2. Since $Ah_T = Ah - \sum_{j \geq 2} Ah_{T_j}$, we have

$$(1 - \delta_{2s})\|h_T\|_2^2 \leq \|Ah_T\|_{\ell_2}^2 = <Ah_T, Ah> - <Ah_T, \sum_{j \geq 2} Ah_{T_j}>$$

Lemma 1.3 gives

$$|<Ah_{T_i}, Ah_{T_j}>| \leq \delta_{2s} \|Ah_T\|_{\ell_2} \|Ah\|_{\ell_2}$$

Note that $\|h_{T_0}\|_2 + \|h_{T_1}\|_2 \leq \sqrt{2}\|h_T\|_2$, we have

$$|<Ah_T, \sum_{j \geq 2} Ah_{T_j}>| = |\sum_{j \geq 2} <Ah_{T_0}, Ah_{T_j}> + \sum_{j \geq 2} <Ah_{T_1}, Ah_{T_j}>|$$

$$\leq \delta_{2s}(\|h_{T_0}\|_2 + \|h_{T_1}\|_2) \sum_{j \geq 2} \|h_{T_j}\|_2 \leq \sqrt{2}\delta_{2s}\|h_T\|_2 \sum_{j \geq 2} \|h_{T_j}\|_2$$

$$\leq \sqrt{2}\delta_{2s}\|h_T\|_2 \frac{\|u_{T_0^c}\|_1}{\sqrt{s}}$$
Preliminaries: Lemma 4

Let $A$ satisfies the RIP with order $2s$ with $\delta_{2s} < \sqrt{2} - 1$. Let $x, \hat{x}$ be given and define $h = \hat{x} - x$. Let $T_0$ denote the index set corresponding to the $s$ entries of $x$ with largest magnitude. Define $T_1$ be the index set corresponding to the $s$ entries of $h_{T_0^c}$. Set $T = T_0 \cup T_1$. If $\|\hat{x}\|_1 \leq \|x\|_1$. Then

$$\|h\|_2 \leq C_0 \frac{\|x - x_s\|_1}{\sqrt{s}} + C_1 \frac{|\langle Ah_T, Ah \rangle|}{\|h_T\|_2}$$

where $C_0 = 2^{1 - (1 - \sqrt{2})\delta_{2s}}$ and $C_1 = \frac{2}{1 - (1 + \sqrt{2})\delta_{2s}}$

**Proof:** Note that $h = h_T + h_{T^c}$, then $\|h\|_2 \leq \|h_T\|_2 + \|h_{T^c}\|_2$. Let $T_j$ be defined similarly as Lemma 2, then we have

$$\|h_{T^c}\|_2 = \| \sum_{j \geq 2} h_{T_j} \|_2 \leq \sum_{j \geq 2} \|h_{T_j}\|_2 \leq \frac{\|h_{T_0^c}\|_1}{\sqrt{s}}$$
Continue: Proof Lemma 4

Since $\|\hat{x}\|_1 \leq \|x\|_1$, we obtain

$$\|x\|_1 \geq \|x_{T_0} + h_{T_0}\|_1 + \|x_{T_0} + h_{T_0}\|_1 \geq \|x_{T_0}\|_1 - \|h_{T_0}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1.$$

Rearranging and again applying the triangle inequality

$$\|h_{T_0}c\|_1 \leq \|x\|_1 - \|x_{T_0}\|_1 + \|h_{T_0}\|_1 + \|x_{T_0}\|_1 \leq \|x - x_{T_0}\|_1 + \|h_{T_0}\|_1 + \|x_{T_0}\|_1.$$

Hence, we have $\|h_{T_0}c\|_1 \leq \|h_{T_0}\|_1 + 2\|x - x_s\|_1$. Therefore,

$$\|h_{T_0}c\|_2 \leq \frac{\|h_{T_0}\|_1 + 2\|x - x_s\|_1}{\sqrt{s}} \leq \|h_{T_0}\|_2 + \frac{2\|x - x_s\|_1}{\sqrt{s}}.$$

Since $\|h_{T_0}\|_2 \leq \|h_T\|_2$, we have

$$\|h\|_2 \leq 2\|h_T\|_2 + \frac{2\|x - x_s\|_1}{\sqrt{s}}$$
Lemma 3 gives

\[ \|h_T\|_2 \leq \alpha \frac{\|h_{T0}\|_1}{\sqrt{s}} + \beta \frac{|\langle Ah_T, Ah \rangle|}{\|h_T\|_2} \]

\leq \alpha \frac{\|h_{T0}\|_1 + 2\|x - x_s\|_1}{\sqrt{s}} + \beta \frac{|\langle Ah_T, Ah \rangle|}{\|h_T\|_2}

\leq \alpha \|h_{T0}\|_2 + 2\alpha \frac{\|x - x_s\|_1}{\sqrt{s}} + \beta \frac{|\langle Ah_T, Ah \rangle|}{\|h_T\|_2}

Using \(\|h_{T0}\|_2 \leq \|h_T\|_2\) gives

\[ (1 - \alpha)\|h_T\|_2 \leq 2\alpha \frac{\|x - x_s\|_1}{\sqrt{s}} + \beta \frac{|\langle Ah_T, Ah \rangle|}{\|h_T\|_2}. \]

Dividing by \(1 - \alpha\) gives

\[ \|h\|_2 \leq \left( \frac{4\alpha}{1 - \alpha} + 2 \right) \frac{\|x - x_s\|_1}{\sqrt{s}} + \frac{2\beta}{1 - \alpha} \frac{|\langle Ah_T, Ah \rangle|}{\|h_T\|_2}. \]
Proof of noisy recovery result

Since \( \hat{x} \) and \( x \) are feasible, we obtain

\[
\|Ah\|_{\ell_2} \leq \|A\hat{x} - b\|_{\ell_2} + \|b - Ax\|_{\ell_2} \leq 2\epsilon
\]

The RIP gives

\[
|\langle Ah_T, Ah \rangle| \leq \|Ah_T\|_2 \|Ah\|_2 \leq 2\epsilon \sqrt{1 + \delta_{2s}} \|h_T\|_2.
\]

Hence,

\[
\|h\|_2 \leq C_0 \frac{\|x - x_s\|_1}{\sqrt{s}} + C_1 \frac{|\langle Ah_T, Ah \rangle|}{\|h_T\|_2}
\]

\[
\leq C_0 \frac{\|x - x_s\|_1}{\sqrt{s}} + C_1 2\epsilon \sqrt{1 + \delta_{2s}}
\]