ECE 18-898G: Special Topics in Signal Processing: Sparsity, Structure, and Inference

Sparse Recovery using L1 minimization

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2018

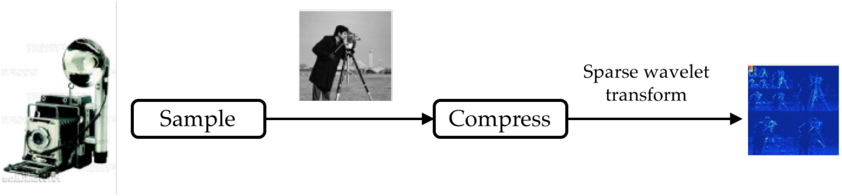
Outline

- ℓ_1 minimization for sparse recovery
- Restricted isometry property (RIP)
- A RIPless theory

Motivation of Compressed Sensing

Conventional paradigms for data acquisition:

- Measure full data
- Compress (by discarding a large fraction of coefficients)



Problem: data is often highly compressible

- Most of acquired data can be thrown away without any perceptual loss

Blind sensing

Ideally, if we know *a priori* which coefficients are worth estimating, then we can simply measure these coefficients

- Unfortunately, we often have no idea which coefficients are most relevant

Compressed sensing: compression on the fly

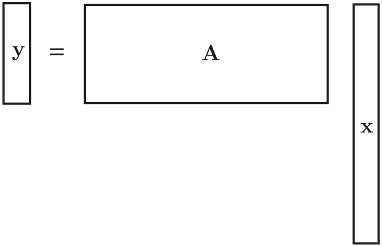
- mimic the behavior of the above ideal situation without pre-computing all coefficients
- often achieved by *random* sensing mechanism

Why go to so much effort to acquire all the data when most of what we get will be thrown away?

Can't we just directly measure the part that won't end up being thrown away?

— David Donoho

Setup: sparse recovery



Recover $\boldsymbol{x} \in \mathbb{R}^p$ given $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n]^\top \in \mathbb{R}^{n \times p}$ $(n \ll p)$: sampling matrix;
$\boldsymbol{a}_i$: sampling vector; $\boldsymbol{x}$: sparse signal

Restricted isometry properties

Optimality for ℓ_0 minimization

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^p} \ \|\boldsymbol{x}\|_0 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

If instead $\exists$ a sparser feasible $\tilde{\boldsymbol{x}} \neq \boldsymbol{x}$ s.t. $\|\tilde{\boldsymbol{x}}\|_0 \leq \|\boldsymbol{x}\|_0 = k$, then

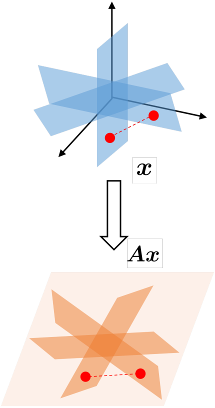
$$\boldsymbol{A}\left(\boldsymbol{x} - \tilde{\boldsymbol{x}}\right) = \boldsymbol{0}. \tag{6.1}$$

We don't want (6.1) to happen, so we hope

$$\boldsymbol{A}(\underbrace{\boldsymbol{x} - \tilde{\boldsymbol{x}}}_{2k-\text{sparse}}) \neq \boldsymbol{0}, \qquad \forall \tilde{\boldsymbol{x}} \ \text{ with } \|\tilde{\boldsymbol{x}}\|_0 \leq k$$

To simultaneously account for all k-sparse $\boldsymbol{x}$, we hope $\boldsymbol{A}_T$ $(|T| \leq 2k)$ to have full column rank, where $\boldsymbol{A}_T$ consists of all columns of $\boldsymbol{A}$ at indices from T

Restricted isometry property (RIP)



Definition 6.1 (Restricted isometry constant)

Restricted isometry constant δ_k of $\boldsymbol{A}$ is smallest quantity s.t.

$$(1-\delta_k)\|\boldsymbol{x}\|^2 \le \|\boldsymbol{A}\boldsymbol{x}\|^2 \le (1+\delta_k)\|\boldsymbol{x}\|^2 \qquad (6.2)$$

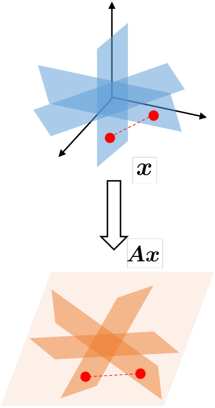
holds for all k-sparse vector $\boldsymbol{x} \in \mathbb{R}^p$

- Equivalently, (6.2) says

$$\max_{S:|S|=k} \underbrace{\|\boldsymbol{A}_S^\top \boldsymbol{A}_S - \boldsymbol{I}_k\|}_{\text{near orthonormality}} = \delta_k$$

where $\boldsymbol{A}_S$ consists of all columns of $\boldsymbol{A}$ at indices from S

Restricted isometry property (RIP)



Definition 6.1 (Restricted isometry constant)

Restricted isometry constant δ_k of $\boldsymbol{A}$ is smallest quantity s.t.

$$(1-\delta_k)\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|^2 \leq (1+\delta_k)\|\boldsymbol{x}\|^2 \tag{6.2}$$

holds for all k-sparse vector $\boldsymbol{x} \in \mathbb{R}^p$

- (Homework) For any $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ that are supported on disjoint subsets S_1, S_2 with $|S_1| \leq s_1$ and $|S_2| \leq s_2$:

$$|\langle \boldsymbol{A}\boldsymbol{x}_1, \boldsymbol{A}\boldsymbol{x}_2 \rangle| \leq \delta_{s_1+s_2}\|\boldsymbol{x}_1\|_2\|\boldsymbol{x}_2\|_2 \tag{6.3}$$

angle-preserving! (consequence of parallelogram identity)

RIP and ℓ_0 minimization

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^p} \ \|\boldsymbol{x}\|_0 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

Fact 6.2

Suppose x is k-sparse. If $\delta_{2k} < 1$, then ℓ_0-minimization is exact and unique.

RIP and ℓ_1 minimization

Theorem 6.3 (Candès 2008)

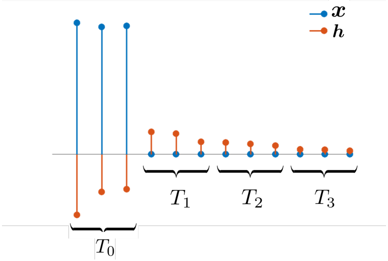
Suppose $\boldsymbol{x}$ is k-sparse. If $\delta_{2k} < \sqrt{2} - 1$, then ℓ_1-minimization is exact and unique.

- RIP implies success of ℓ_1 minimization (also many other methods, as we'll see from later lectures)
- A universal result: works simultaneously for **all** k-sparse signals
- As we will see later, many random designs satisfy this condition with *near-optimal sample complexity*

$$m \sim O(k \log(n/k))$$

Proof of Theorem 6.3

Suppose $\boldsymbol{x}+\boldsymbol{h}$ is feasible and obeys $\|\boldsymbol{x}+\boldsymbol{h}\|_1 \leq \|\boldsymbol{x}\|_1$. The goal is to show that $\boldsymbol{h}=\boldsymbol{0}$ under RIP.



The key is to decompose $\boldsymbol{h}$ into $\boldsymbol{h}_{T_0}+\boldsymbol{h}_{T_1}+\dots$

- T_0: locations of k largest entries of $\boldsymbol{x}$
- T_1: locations of k largest entries of $\boldsymbol{h}$ in $T_0{}^{\mathsf{c}}$
- T_2: locations of k largest entries of $\boldsymbol{h}$ in $(T_0 \cup T_1)^{\mathsf{c}}$
- ...

Proof of Theorem 6.3

The proof proceeds by showing that

1. $\boldsymbol{h}_{T_0 \cup T_1}$ dominates $\boldsymbol{h}_{(T_0 \cup T_1)^{\mathsf{c}}}$ (by objective function)
2. (converse) $\boldsymbol{h}_{(T_0 \cup T_1)^{\mathsf{c}}}$ dominates $\boldsymbol{h}_{T_0 \cup T_1}$ (by RIP + feasibility)

These can happen simultaneously only when $\boldsymbol{h}$ vanishes

Proof of Theorem 6.3

Step 1 (depending only on objective function). Show that

$$\sum\nolimits_{j\geq 2} \|\boldsymbol{h}_{T_j}\| \leq \frac{1}{\sqrt{k}} \|\boldsymbol{h}_{T_0}\|_1. \tag{6.4}$$

This follows immediately by combining the following 2 observations:

(i) Since $\boldsymbol{x} + \boldsymbol{h}$ is assumed to be a better estimate:

$$\|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}+\boldsymbol{h}\|_1 = \underbrace{\|\boldsymbol{x}+\boldsymbol{h}_{T_0}\|_1 + \|\boldsymbol{h}_{T_0^{\mathsf{c}}}\|_1}_{\text{since } T_0 \text{ is support of } \boldsymbol{x}} \geq \underbrace{\|\boldsymbol{x}\|_1 - \|\boldsymbol{h}_{T_0}\|_1}_{\text{triangle inequality}} + \|\boldsymbol{h}_{T_0^{\mathsf{c}}}\|_1$$

$$\Longrightarrow \quad \|\boldsymbol{h}_{T_0^{\mathsf{c}}}\|_1 \leq \|\boldsymbol{h}_{T_0}\|_1 \tag{6.5}$$

(ii) Since entries of $\boldsymbol{h}_{T_{j-1}}$ uniformly dominate those of $\boldsymbol{h}_{T_j}$ $(j \geq 2)$:

$$\|\boldsymbol{h}_{T_j}\| \leq \sqrt{k}\|\boldsymbol{h}_{T_j}\|_\infty \leq \sqrt{k}\frac{\|\boldsymbol{h}_{T_{j-1}}\|_1}{k} = \frac{1}{\sqrt{k}}\|\boldsymbol{h}_{T_{j-1}}\|_1$$

$$\Longrightarrow \quad \sum_{j\geq 2}\|\boldsymbol{h}_{T_j}\| \leq \frac{1}{\sqrt{k}}\sum_{j\geq 2}\|\boldsymbol{h}_{T_{j-1}}\|_1 = \frac{1}{\sqrt{k}}\|\boldsymbol{h}_{T_0^{\mathsf{c}}}\|_1 \tag{6.6}$$

Proof of Theorem 6.3

Step 2 (using feasibility + RIP). Show that $\exists \rho < 1$ s.t.

$$\|\boldsymbol{h}_{T_0 \cup T_1}\| \leq \rho \sum\nolimits_{j \geq 2} \|\boldsymbol{h}_{T_j}\| \tag{6.7}$$

If this claim holds, then

$$\begin{aligned} \|\boldsymbol{h}_{T_0 \cup T_1}\| \leq \rho \sum\nolimits_{j \geq 2} \|\boldsymbol{h}_{T_j}\| &\overset{(6.4)}{\leq} \rho \frac{1}{\sqrt{k}} \|\boldsymbol{h}_{T_0}\|_1 \\ &\leq \rho \frac{1}{\sqrt{k}} \left(\sqrt{k} \|\boldsymbol{h}_{T_0}\| \right) = \rho \|\boldsymbol{h}_{T_0}\| \ \leq \ \rho \|\boldsymbol{h}_{T_0 \cup T_1}\|. \end{aligned} \tag{6.8}$$

Since $\rho < 1$, we necessarily have $\boldsymbol{h}_{T_0 \cup T_1} = \boldsymbol{0}$, which together with (6.5) yields $\boldsymbol{h} = \boldsymbol{0}$

Proof of Theorem 6.3

We now prove (6.7). To connect $\boldsymbol{h}_{T_0 \cup T_1}$ with $\boldsymbol{h}_{(T_0 \cup T_1)^c}$, we use feasibility:

$$\boldsymbol{A}\boldsymbol{h} = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1} = -\sum\nolimits_{j \geq 2} \boldsymbol{A}\boldsymbol{h}_{T_j},$$

which taken collectively with RIP yields

$$(1-\delta_{2k})\|\boldsymbol{h}_{T_0 \cup T_1}\|^2 \leq \|\boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1}\|^2 = \left|\langle \boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1}, \sum\nolimits_{j \geq 2} \boldsymbol{A}\boldsymbol{h}_{T_j}\rangle\right|.$$

It follows from (6.3) that for all $j \geq 2$,

$$\begin{aligned} |\langle \boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1}, \boldsymbol{A}\boldsymbol{h}_{T_j}\rangle| &\leq |\langle \boldsymbol{A}\boldsymbol{h}_{T_0}, \boldsymbol{A}\boldsymbol{h}_{T_j}\rangle| + |\langle \boldsymbol{A}\boldsymbol{h}_{T_1}, \boldsymbol{A}\boldsymbol{h}_{T_j}\rangle| \\ &\overset{(6.3)}{\leq} \delta_{2k}(\|\boldsymbol{h}_{T_0}\| + \|\boldsymbol{h}_{T_1}\|)\|\boldsymbol{h}_{T_j}\| \leq \delta_{2k}\sqrt{2}\|\boldsymbol{h}_{T_0 \cup T_1}\| \cdot \|\boldsymbol{h}_{T_j}\|, \end{aligned}$$

which gives

$$\begin{aligned} (1-\delta_{2k})\|\boldsymbol{h}_{T_0 \cup T_1}\|^2 &\leq \sum\nolimits_{j \geq 2} |\langle \boldsymbol{A}\boldsymbol{h}_{T_0 \cup T_1}, \boldsymbol{A}\boldsymbol{h}_{T_j}\rangle| \\ &\leq \sqrt{2}\delta_{2k}\|\boldsymbol{h}_{T_0 \cup T_1}\| \sum\nolimits_{j \geq 2} \|\boldsymbol{h}_{T_j}\| \end{aligned}$$

This establishes (6.7) if $\rho := \frac{\sqrt{2}\delta_{2k}}{1-\delta_{2k}} < 1$ (or equivalently, $\delta_{2k} < \sqrt{2} - 1$).

Robustness for compressible signals

Theorem 6.4

If $\delta_{2k} < \sqrt{2} - 1$, then the solution $\hat{\boldsymbol{x}}$ to ℓ_1-minimization obeys

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\| \lesssim \frac{\|\boldsymbol{x} - \boldsymbol{x}_k\|_1}{\sqrt{k}},$$

where $\boldsymbol{x}_k$ is best k-term approximation of $\boldsymbol{x}$

- Suppose l^{th} largest entry of $\boldsymbol{x}$ is $1/l^{\alpha}$ for some $\alpha > 1$, then

$$\frac{1}{\sqrt{k}} \|\boldsymbol{x} - \boldsymbol{x}_k\|_1 \approx \frac{1}{\sqrt{k}} \sum_{l>k} l^{-\alpha} \approx k^{-\alpha+0.5} \ll 1$$

- ℓ_1-min works well in recovering compressible signals
- Follows similar arguments as in proof of Theorem 6.3

Proof of Theorem 6.4

Step 1 (depending only on objective function). Show that

$$\sum_{j\geq 2} \|\boldsymbol{h}_{T_j}\| \leq \frac{1}{\sqrt{k}} \|\boldsymbol{h}_{T_0}\|_1 + \frac{2}{\sqrt{k}} \|\boldsymbol{x} - \boldsymbol{x}_{T_0}\|_1. \tag{6.9}$$

This follows immediately by combining the following 2 observations:

(i) Since $\boldsymbol{x} + \boldsymbol{h}$ is assumed to be a better estimate:

$$\begin{aligned} \|\boldsymbol{x}_{T_0}\|_1 + \|\boldsymbol{x}_{T_0^c}\|_1 = \|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x} + \boldsymbol{h}\|_1 &= \|\boldsymbol{x}_{T_0} + \boldsymbol{h}_{T_0}\|_1 + \|\boldsymbol{x}_{T_0^c} + \boldsymbol{h}_{T_0^c}\|_1 \\ &\geq \|\boldsymbol{x}_{T_0}\|_1 - \|\boldsymbol{h}_{T_0}\|_1 + \|\boldsymbol{h}_{T_0^c}\|_1 - \|\boldsymbol{x}_{T_0^c}\|_1 \end{aligned}$$

$$\Longrightarrow \quad \|\boldsymbol{h}_{T_0^c}\|_1 \leq \|\boldsymbol{h}_{T_0}\|_1 + 2\|\boldsymbol{x}_{T_0^c}\|_1 \tag{6.10}$$

(ii) Recall from (6.6) that $\sum_{j\geq 2} \|\boldsymbol{h}_{T_j}\| \leq \frac{1}{\sqrt{k}} \|\boldsymbol{h}_{T_0^c}\|_1$.

We highlight in red the part different from the proof of Theorem 6.3.

Proof of Theorem 6.4

Step 2 (using feasibility + RIP). Recall from (6.7) that $\exists \rho < 1$ s.t.

$$\|\boldsymbol{h}_{T_0 \cup T_1}\| \le \rho \sum_{j \ge 2} \|\boldsymbol{h}_{T_j}\| \tag{6.11}$$

If this claim holds, then

$$\begin{aligned}
\|\boldsymbol{h}_{T_0 \cup T_1}\| &\le \rho \sum_{j \ge 2} \|\boldsymbol{h}_{T_j}\| \overset{(6.10) \text{ and } (6.6)}{\le} \rho \frac{1}{\sqrt{k}} \{\|\boldsymbol{h}_{T_0}\|_1 + 2\|\boldsymbol{x}_{T_0^{\mathsf{c}}}\|_1\} \\
&\le \rho \frac{1}{\sqrt{k}} \left(\sqrt{k}\|\boldsymbol{h}_{T_0}\| + 2\|\boldsymbol{x}_{T_0^{\mathsf{c}}}\|_1\right) = \rho\|\boldsymbol{h}_{T_0}\| + \frac{2\rho}{\sqrt{k}}\|\boldsymbol{x}_{T_0^{\mathsf{c}}}\|_1 \\
&\le \rho\|\boldsymbol{h}_{T_0 \cup T_1}\| + \frac{2\rho}{\sqrt{k}}\|\boldsymbol{x}_{T_0^{\mathsf{c}}}\|_1.
\end{aligned}$$

$$\Longrightarrow \quad \|\boldsymbol{h}_{T_0 \cup T_1}\| \le \frac{2\rho}{1-\rho} \frac{\|\boldsymbol{x}_{T_0^{\mathsf{c}}}\|_1}{\sqrt{k}}. \tag{6.12}$$

We highlight in red the part different from the proof of Theorem 6.3.

Proof of Theorem 6.4

Finally, putting the above together yields

$$
\begin{aligned}
\|\boldsymbol{h}\| &\leq \|\boldsymbol{h}_{T_0\cup T_1}\| + \|\boldsymbol{h}_{(T_0\cup T_1)^{\mathsf{c}}}\| \\
&\overset{(6.9)}{\leq} \|\boldsymbol{h}_{T_0\cup T_1}\| + \frac{1}{\sqrt{k}}\|\boldsymbol{h}_{T_0}\|_1 + \frac{2}{\sqrt{k}}\|\boldsymbol{x}-\boldsymbol{x}_{T_0}\|_1 \\
&\leq \|\boldsymbol{h}_{T_0\cup T_1}\| + \|\boldsymbol{h}_{T_0}\| + \frac{2}{\sqrt{k}}\|\boldsymbol{x}-\boldsymbol{x}_{T_0}\|_1 \\
&\leq 2\|\boldsymbol{h}_{T_0\cup T_1}\| + \frac{2}{\sqrt{k}}\|\boldsymbol{x}-\boldsymbol{x}_{T_0}\|_1 \\
&\overset{(6.12)}{\leq} \frac{2(1+\rho)}{1-\rho}\frac{\|\boldsymbol{x}-\boldsymbol{x}_{T_0}\|_1}{\sqrt{k}}
\end{aligned}
$$

We highlight in red the part different from the proof of Theorem 6.3.

ℓ_1 recovery in the noisy case

In the presence of additive measurement noise,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w},$$

where $\|\boldsymbol{w}\|_2 \leq \epsilon$ is assumed to be bounded.
We can modify the BP algorithm in the following manner:

$$\text{(BP-noisy:)} \quad \hat{\boldsymbol{x}} = \text{argmin}_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

Theorem 6.5 (Performance of BP via RIP, noisy case)

If $\delta_{2k} < \sqrt{2} - 1$, then for any vector $\boldsymbol{x}$, the solution to basis pursuit (noisy case) satisfies

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq C_0 k^{-1/2} \|\boldsymbol{x} - \boldsymbol{x}_k\|_1 + C_1 \epsilon.$$

where $\boldsymbol{x}_k$ is the best k-term approximation of $\boldsymbol{x}$ for some constants C_0 and C_1.

Proof of Theorem 6.5

Again let's start by assuming $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{h}$. The key difference from the noiseless case is that in Step 2, we now have

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{h}\|_2 = \|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x})\|_2 &= \|(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}) - (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})\|_2 \\ &\le \|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_2 + \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \le 2\epsilon.\end{aligned}$$

Therefore, we need to bound

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{h}_{T_0\cup T_1}\|_2^2 &= \langle \boldsymbol{A}\boldsymbol{h} - \sum_{j\ge 2} \boldsymbol{A}\boldsymbol{h}_{T_j}, \boldsymbol{A}\boldsymbol{h}_{T_0\cup T_1}\rangle \\ &\le \underbrace{\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{h}_{T_0\cup T_1}\rangle}_{\le 2\epsilon\delta_{2k}\|\boldsymbol{h}_{T_0\cup T_1}\|_2} \underbrace{- \sum_{j\ge 2}\langle \boldsymbol{A}\boldsymbol{h}_{T_j}, \boldsymbol{A}\boldsymbol{h}_{T_0\cup T_1}\rangle}_{\text{bounded as before}}\end{aligned}$$

By plugging in this modification, we show

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 = \|\boldsymbol{h}\|_2 \le \frac{2(1+\rho)}{1-\rho}\frac{\|\boldsymbol{x} - \boldsymbol{x}_k\|_1}{\sqrt{k}} + \frac{2\alpha}{1-\rho}\epsilon,$$

where $\alpha = \frac{2\sqrt{1+\delta_{2k}}}{1-\delta_{2k}}$.

Which design matrix satisfies RIP?

First example: i.i.d. Gaussian design

Lemma 6.6

A random matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ *with i.i.d.* $\mathcal{N}\left(0, \frac{1}{n}\right)$ *entries satisfies* $\delta_k < \delta$ *with high prob., as long as*

$$n \gtrsim \frac{1}{\delta^2} k \log \frac{p}{k}$$

- This is where non-asymptotic random matrix theory enters

Gaussian random matrices

Lemma 6.7 (See Vershynin '10)

Suppose $\boldsymbol{B} \in \mathbb{R}^{n\times k}$ *is composed of i.i.d.* $\mathcal{N}(0,1)$ *entries. Then*

$$\begin{cases} \mathbb{P}\Big(\frac{1}{\sqrt{n}}\sigma_{\max}(\boldsymbol{B}) > 1+\sqrt{\frac{k}{n}}+t\Big) & \leq e^{-nt^2/2} \\ \mathbb{P}\Big(\frac{1}{\sqrt{n}}\sigma_{\min}(\boldsymbol{B}) < 1-\sqrt{\frac{k}{n}}-t\Big) & \leq e^{-nt^2/2}. \end{cases}$$

- When $n \gg k$, one has $\frac{1}{n}\boldsymbol{B}^\top \boldsymbol{B} \approx \boldsymbol{I}_k$
- Similar results (up to different constants) hold for i.i.d. sub-Gaussian matrix

Proof of Lemma 6.6

1. Fix any index subset $S \subseteq \{1, \cdots, \}$, $|S| = k$, then $\boldsymbol{A}_S$ (submatrix of $\boldsymbol{A}$ consisting of columns at indices from S) obeys

$$\left\| \boldsymbol{A}_S^\top \boldsymbol{A}_S - \boldsymbol{I}_k \right\| \leq O\Big(\sqrt{k/n}\Big) + t$$

with prob. exceeding $1 - 2e^{-c_1 n t^2}$, where $c_1 > 0$ is constant.

2. Taking a union bound over all $S \subseteq \{1, \cdots, p\}$, $|S| = k$ yields

$$\delta_k = \max_{S:|S|=k} \left\| \boldsymbol{A}_S^\top \boldsymbol{A}_S - \boldsymbol{I}_k \right\| \leq O\Big(\sqrt{k/n}\Big) + t$$

with prob. exceeding $1 - 2\binom{p}{k} e^{-c_1 n t^2} \geq 1 - 2e^{k \log(ep/k) - c_1 n t^2}$. Thus, $\delta_k < \delta$ with high prob. as long as $n \gtrsim \delta^{-2} k \log(p/k)$.

Other design matrices that satisfy RIP

- Random matrices with i.i.d. sub-Gaussian entries, as long as

$$n \gtrsim k \log(p/k)$$

- Random partial DFT matrices with

$$n \gtrsim k \log^4 p,$$

  where rows of $\boldsymbol{A}$ are independently sampled from rows of DFT matrix $\boldsymbol{F}$ (Rudelson & Vershynin '08)

  - If you have learned entropy method / generic chaining, check out Rudelson & Vershynin '08 and Candes & Plan '11

Other design matrices that satisfy RIP

- Random convolution matrices with

$$n \gtrsim k \log^4 p,$$

where rows of $\boldsymbol{A}$ are independently sampled from rows of

$$\boldsymbol{G} = \begin{bmatrix} g_0 & g_1 & g_2 & \cdots & g_{p-1} \\ g_{p-1} & g_0 & g_1 & \cdots & g_{p-2} \\ g_{p-2} & g_{p-1} & g_0 & \cdots & g_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1 & g_2 & g_3 & \cdots & g_0 \end{bmatrix}$$

with $\mathbb{P}(g_i = \pm 1) = 0.5$ (Krahmer, Mendelson, & Rauhut '14)

A RIPless theory

Is RIP necessary?

- RIP leads to a universal result holding simultaneously for all k-sparse $\boldsymbol{x}$
  - Universality is often not needed as we might only care about a particular $\boldsymbol{x}$

- There may be a gap between the regime where RIP holds and the regime in which one has minimal measurements

- Certifying RIP is hard

Can we develop a non-universal RIPless theory?

A standard recipe

1. Write out Karush-Kuhn-Tucker (KKT) optimality conditions
   - typically involve certain dual variables

2. Construct dual variables satisfying KKT conditions

Karush-Kuhn-Tucker (KKT) condition

Consider a convex problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\nu}) := f(\boldsymbol{x}) + \boldsymbol{\nu}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) \qquad (\boldsymbol{\nu} : \text{Lagrangian multiplier})$$

If $\boldsymbol{x}$ is optimizer, then KKT optimality condition reads

$$\begin{cases} \boldsymbol{0} = \nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{v}) \\ \boldsymbol{0} \in \partial_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{v}) \end{cases}$$

Karush-Kuhn-Tucker (KKT) condition

Consider a convex problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\nu}) := f(\boldsymbol{x}) + \boldsymbol{\nu}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) \qquad (\boldsymbol{\nu} : \text{Lagrangian multiplier})$$

If $\boldsymbol{x}$ is optimizer, then KKT optimality condition reads

$$\begin{cases} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \\ \boldsymbol{0} \in \partial f(\boldsymbol{x}) + \boldsymbol{A}^\top \boldsymbol{\nu} \quad (\text{no constraint on } \boldsymbol{\nu}) \end{cases}$$

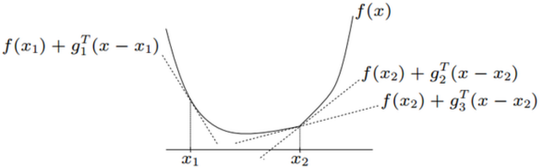
Subgradient

Consider a convex function $f(\boldsymbol{x})$ (possibly nonsmooth).

Definition 6.8 (Subgradient)

$\boldsymbol{u} \in \partial f(\boldsymbol{x}_0)$ is a subgradient of a convex f at $\boldsymbol{x}_0$ if for all $\boldsymbol{x}$:

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \boldsymbol{u}^T(\boldsymbol{x} - \boldsymbol{x}_0)$$



Remark: if f is differentiable at $\boldsymbol{x}_0$, the only subgradient is the gradient $\nabla f(\boldsymbol{x}_0)$.

Subgradient of ℓ_1 norm

Example: For the scalar absolute function $f(t) = |t|$, $t \in \mathbb{R}$, $u \in \partial f(t)$ iff

$$\begin{cases} u = \mathsf{sgn}(t), & t \neq 0 \\ u \in [-1, 1], & t = 0 \end{cases}$$

Example: For $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{u} \in \partial f(\boldsymbol{x})$ iff

$$\begin{cases} u_i = \mathsf{sgn}(x_i), & x_i \neq 0 \\ u_i \in [-1, 1], & x_i = 0 \end{cases}$$

KKT condition for ℓ_1 minimization

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & \|\boldsymbol{x}\|_1 \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \end{aligned}$$

If $\boldsymbol{x}$ is optimizer, then KKT optimality condition reads

$$\begin{cases} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0}, & \text{(naturally satisfied as } \boldsymbol{x} \text{ is truth)} \\ \boldsymbol{0} \in \partial\|\boldsymbol{x}\|_1 + \boldsymbol{A}^\top \boldsymbol{\nu} & \text{(no constraint on } \boldsymbol{\nu}\text{)} \end{cases}$$

$$\Longleftrightarrow \quad \exists \boldsymbol{u} \in \text{range}(\boldsymbol{A}^\top) \quad \text{s.t.} \quad \underbrace{\begin{cases} u_i = \text{sign}(x_i), & \text{if } x_i \neq 0 \\ u_i \in [-1, 1], & \text{else} \end{cases}}_{\boldsymbol{u} \text{ is a valid subgradient}}$$

Depends only on signs of x_i's irrespective of their magnitudes

Uniqueness

Theorem 6.9 (A sufficient—and almost necessary—condition)

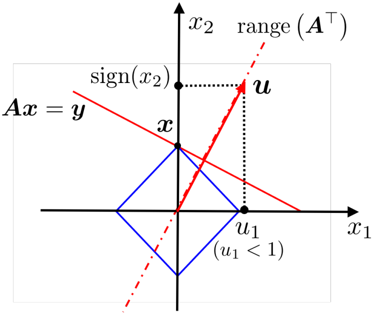
Let $T := \mathrm{supp}(\boldsymbol{x})$. Suppose $\boldsymbol{A}_T$ has full column rank. If

$$\exists \boldsymbol{u} = \boldsymbol{A}^\top \boldsymbol{\nu} \text{ for some } \boldsymbol{\nu} \in \mathbb{R}^n \quad \text{s.t.} \quad \begin{cases} u_i & = \mathrm{sign}(x_i), \quad \text{if } x_i \neq 0 \\ u_i & \in (-1, 1), \quad \text{else} \end{cases},$$

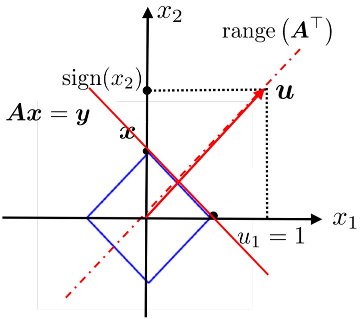
then $\boldsymbol{x}$ is unique solution to ℓ_1 minimization.

- Only slightly stronger than KKT!
- $\boldsymbol{\nu}$ is said to be a dual certificate
  - recall that $\boldsymbol{\nu}$ is Lagrangian multiplier
- *Finding $\boldsymbol{\nu}$ comes down to solving another convex problem*

Geometric interpretation of dual certificate



When $|u_1| < 1$, solution is unique



When $|u_1| = 1$, solution is non-unique

When we are able to find $\boldsymbol{u} \in \mathrm{range}(\boldsymbol{A}^\top)$ s.t. $u_2 = \mathrm{sign}(x_2)$ and $|u_1| < 1$, then $\boldsymbol{x}$ (with $x_1 = 0$) is unique solution to ℓ_1 minimization

Proof of Theorem 6.9

Let $\boldsymbol{w} \in \partial\|\boldsymbol{x}\|_1$ be $\begin{cases} w_i = \mathrm{sign}(x_i), & \text{if } i \in T \text{ (support of } \boldsymbol{x}\text{)}; \\ w_i = \mathrm{sign}(h_i), & \text{else.} \end{cases}$ If $\boldsymbol{x} + \boldsymbol{h}$ is optimizer with $\boldsymbol{h}_{T^c} \neq \mathbf{0}$, then

$$
\begin{aligned}
\|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}+\boldsymbol{h}\|_1 \quad &\geq \quad \|\boldsymbol{x}\|_1 + \langle \boldsymbol{w}, \boldsymbol{h} \rangle = \|\boldsymbol{x}\|_1 + \langle \boldsymbol{u}, \boldsymbol{h} \rangle + \langle \boldsymbol{w} - \boldsymbol{u}, \boldsymbol{h} \rangle \\
&= \quad \|\boldsymbol{x}\|_1 + \langle \underbrace{\boldsymbol{A}^\top \boldsymbol{\nu}}_{\text{assumption on } \boldsymbol{u}}, \boldsymbol{h} \rangle + \sum_{i \notin T} (\mathrm{sign}(h_i) h_i - u_i h_i) \\
&= \quad \|\boldsymbol{x}\|_1 + \langle \boldsymbol{\nu}, \underbrace{\boldsymbol{A}\boldsymbol{h}}_{:=\mathbf{0} \text{ (feasibility)}} \rangle + \sum_{i \notin T} (|h_i| - u_i h_i) \\
&\geq \quad \|\boldsymbol{x}\|_1 + \sum\nolimits_{i \notin T} (1 - |u_i|)\, |h_i| \; > \; \|\boldsymbol{x}\|_1,
\end{aligned}
$$

resulting in contradiction.

Further, when $\boldsymbol{h}_{T^c} = \mathbf{0}$, one must have $\boldsymbol{h}_T = \mathbf{0}$ from left-invertibility of $\boldsymbol{A}_T$, and hence $\boldsymbol{h} = \boldsymbol{h}_T + \boldsymbol{h}_{T^c} = \mathbf{0}$

Constructing dual certificates under Gaussian design

We illustrate how to construct dual certificates for the following setup

- $\boldsymbol{x} \in \mathbb{R}^p$ is k-sparse
- Entries of $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ are i.i.d. standard Gaussian
- Sample size n obeys

$$n \gtrsim k \log p$$

Constructing dual certificates under Gaussian design

$$\begin{aligned} \text{Find} \quad & \boldsymbol{\nu} \in \mathbb{R}^n \\ \text{s.t.} \quad & (\boldsymbol{A}^\top \boldsymbol{\nu})_T = \mathsf{sign}(\boldsymbol{x}_T) && (6.13) \\ & |(\boldsymbol{A}^\top \boldsymbol{\nu})_i| < 1, \qquad i \notin T && (6.14) \end{aligned}$$

Step 1: propose a $\boldsymbol{\nu}$ compatible with linear constraints (6.13). One candidate is *least squares* solution:

$$\boldsymbol{\nu} = \boldsymbol{A}_T (\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1} \mathsf{sign}(\boldsymbol{x}_T) \qquad \text{(explicit expression)}$$

- LS solution minimizes $\|\boldsymbol{\nu}\|$, which will also be helpful when controlling $|(\boldsymbol{A}^\top \boldsymbol{\nu})_i|$
- From Lemma 6.7, $\boldsymbol{A}_T^\top \boldsymbol{A}_T$ is invertible when $n \gtrsim k \log p$

Constructing dual certificates under Gaussian design

Step 2: verify (6.14), which amounts to controlling

$$\max_{i \notin T} \Big| \Big\langle \underbrace{\boldsymbol{a}_i}_{i\text{th column of } \boldsymbol{A}}, \underbrace{\boldsymbol{A}_T(\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}\mathsf{sign}(\boldsymbol{x}_T)}_{\boldsymbol{\nu}} \Big\rangle \Big|$$

- Since $\boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and $\boldsymbol{\nu}$ are independent for any $i \notin T$,

$$\max_{i \notin T} |\langle \boldsymbol{a}_i, \ \boldsymbol{\nu} \rangle| \lesssim \|\boldsymbol{\nu}\| \sqrt{\log p}$$

- $\|\boldsymbol{\nu}\|$ can be bounded by

$$\begin{aligned} \|\boldsymbol{\nu}\| &\leq \|\boldsymbol{A}_T(\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}\| \cdot \|\mathsf{sgn}(\boldsymbol{x}_T)\| \\ &= \|(\underbrace{\boldsymbol{A}_T^\top \boldsymbol{A}_T}_{\text{eigenvalues } \asymp n})^{-1/2}\| \cdot \sqrt{k} \lesssim \sqrt{k/n} \end{aligned}$$

- When $n/(k \log p)$ is sufficiently large, $\max_{i \notin T} |\langle \boldsymbol{a}_i, \ \boldsymbol{\nu} \rangle| < 1$

Details

- Conditioned on $\boldsymbol{\nu}$, $\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle \sim \mathcal{N}(0, \|\boldsymbol{\nu}\|_2^2)$, we have the Chernoff bound for the tail of a Gaussian rv:

$$\mathbb{P}\left(|\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle| \geq 1 | \boldsymbol{\nu}\right) \leq 2\exp\left(-\frac{1}{2\|\boldsymbol{\nu}\|_2^2}\right)$$

- With probability at least $1 - e^{-cn}$, we could also bound $\|\boldsymbol{\nu}\|_2$ as

$$\|\boldsymbol{\nu}\|_2 \leq \sqrt{\frac{2k}{n}}$$

- We have

$$\begin{aligned}\mathbb{P}(\max_{i\in T^c} |\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle| \geq 1) &\leq |T^c| \cdot \mathbb{P}(|\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle| > 1) \quad \text{union bound} \\ &\leq p \int_{\boldsymbol{\nu}} \mathbb{P}(|\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle| \geq 1 | \boldsymbol{\nu}) d\mu(\boldsymbol{\nu}).\end{aligned}$$

Continued

Note that

$$
\begin{aligned}
&\int_{\boldsymbol{\nu}} \mathbb{P}(|\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle| \geq 1|\boldsymbol{\nu}) d\mu(\boldsymbol{\nu}) \\
&= \left(\int_{\|\boldsymbol{\nu}\|_2 \leq \sqrt{\frac{2k}{n}}} + \int_{\|\boldsymbol{\nu}\|_2 > \sqrt{\frac{2k}{n}}} \right) \mathbb{P}(|\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle| \geq 1|\boldsymbol{\nu}) d\mu(\boldsymbol{\nu}) \\
&\leq \int_{\|\boldsymbol{\nu}\|_2 \leq \sqrt{\frac{2k}{n}}} \mathbb{P}(|\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle| \geq 1|\boldsymbol{\nu}) d\mu(\boldsymbol{\nu}) + \mathbb{P}\left(\|\boldsymbol{\nu}\|_2 > \sqrt{\frac{2k}{n}} \right) \\
&\leq \int_{\|\boldsymbol{\nu}\|_2 \leq \sqrt{\frac{2k}{n}}} 2e^{-\frac{1}{2\|\boldsymbol{\nu}\|_2^2}} d\mu(\boldsymbol{\nu}) + e^{-cn} \\
&\leq 2e^{-\frac{n}{4k}} + e^{-cn} \leq 3e^{-\frac{n}{4k}},
\end{aligned}
$$

which gives

$$
\mathbb{P}(\max_{i \in T^c} |\langle \boldsymbol{a}_i,\ \boldsymbol{\nu}\rangle| \geq 1) \leq 3pe^{-\frac{n}{4k}} \leq p^{-\gamma}
$$

by setting $n = 4(\gamma + 1)k \log p$ for some constant $\gamma > 0$.

More general random sampling

Consider a random design: each sampling vector $\boldsymbol{a}_i$ is independently drawn from a distribution F

$$\boldsymbol{a}_i \sim F$$

Incoherence sampling:

- *Isotropy:*

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \boldsymbol{a} \sim F$$

  - components of $\boldsymbol{a}$: (i) unit variance; (ii) uncorrelated

- *Incoherence:* let $\mu(F)$ be the smallest quantity s.t. for $\boldsymbol{a} \sim F$,

$$\|\boldsymbol{a}\|_\infty^2 \leq \mu(F) \quad \text{with high probability}$$

Incoherence

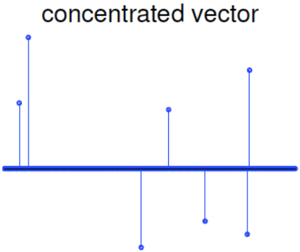
We want $\mu(F)$ (resp. $\boldsymbol{A}$) to be small (resp. dense)!

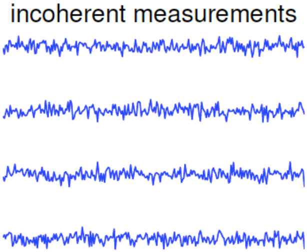
What happen if sampling vectors $\boldsymbol{a}_i$ are sparse?

- Example: $\boldsymbol{a}_i \sim$ Uniform$(\{\sqrt{p}\boldsymbol{e}_1, \cdots, \sqrt{p}\boldsymbol{e}_p\})$

$$\underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{\substack{\boldsymbol{y} \\ \text{no information}}} = \underbrace{\sqrt{p}\begin{bmatrix} 1 & & & & & & & & & \\ & & & & & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & & & & 1 \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ 0 \\ 0 \\ 5 \\ 0 \\ 0 \end{bmatrix}}_{\boldsymbol{x}}$$

Incoherent random sampling





A general RIPless theory

Theorem 6.10 (Candes & Plan, '11)

Suppose $\boldsymbol{x} \in \mathbb{R}^p$ *is* k*-sparse, and* $\boldsymbol{a}_i \overset{\text{ind.}}{\sim} F$ *is isotropic. Then* ℓ_1 *minimization is exact and unique with high prob., provided that*

$$n \gtrsim \mu(F) k \log p$$

- Near-optimal even for highly structured sampling matrices
- Proof idea: produce an (approximate) dual certificate by a clever *golfing scheme* pioneered by David Gross

Examples of incoherent sampling

- **Binary sensing:** $\mathbb{P}(a[i] = \pm 1) = 0.5$:

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \|\boldsymbol{a}\|_\infty^2 = 1, \qquad \mu = 1$$

$$\Longrightarrow \quad \ell_1\text{-min succeeds if } n \gtrsim k \log p$$

- **Gaussian sensing:** $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$:

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \|\boldsymbol{a}\|_\infty^2 \lesssim 2 \log p \quad \Rightarrow \quad \mu \asymp \log p$$

- **Partial Fourier transform:** pick a random frequency $f \sim \text{Unif}\{0, \frac{1}{p}, \cdots, \frac{p-1}{p}\}$ or $f \sim \text{Unif}[0,1]$ and set $a[i] = e^{j2\pi f i}$:

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \|\boldsymbol{a}\|_\infty^2 = 1, \qquad \mu = 1$$

$$\Longrightarrow \quad \ell_1\text{-min succeeds if } n \gtrsim k \log p$$

  - Improves upon RIP-based result ($n \gtrsim k \log^4 p$)

Examples of incoherent sampling

- Random convolution matrices: rows of $\boldsymbol{A}$ are independently sampled from rows of

$$\boldsymbol{G} = \begin{bmatrix} g_0 & g_1 & g_2 & \cdots & g_{p-1} \\ g_{p-1} & g_0 & g_1 & \cdots & g_{p-2} \\ g_{p-2} & g_{p-1} & g_0 & \cdots & g_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1 & g_2 & g_3 & \cdots & g_0 \end{bmatrix}$$

  with $\mathbb{P}(g_i = \pm 1) = 0.5$. One has

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \boldsymbol{I}, \qquad \|\boldsymbol{a}\|_\infty^2 = 1, \qquad \mu = 1$$

$$\Longrightarrow \quad \ell_1\text{-min succeeds if } n \gtrsim k \log p$$

  - Improves upon RIP-based result ($n \gtrsim k \log^4 p$)

A general scheme for dual construction

$$\begin{aligned} \text{Find} \quad & \boldsymbol{\nu} \in \mathbb{R}^n \\ \text{s.t.} \quad & \boldsymbol{A}_T^\top \boldsymbol{\nu} = \mathsf{sign}(\boldsymbol{x}_T) && (6.15) \\ & \|\boldsymbol{A}_{T^c}^\top \boldsymbol{\nu}\|_\infty < 1 && (6.16) \end{aligned}$$

A candidate: least squares solution w.r.t. (6.19)

$$\boldsymbol{\nu} = \boldsymbol{A}_T (\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1} \mathsf{sign}(\boldsymbol{x}_T) \qquad \text{(explicit expression)}$$

To verify (6.20), we need to control $\boldsymbol{A}_{T^c}^\top \boldsymbol{A}_T (\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1} \mathsf{sign}(\boldsymbol{x}_T)$

- Issue 1: in general, $\boldsymbol{A}_{T^c}$ and $\boldsymbol{A}_T$ are dependent
- Issue 2: $(\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}$ is hard to deal with

A general scheme for dual construction

$$\begin{aligned} \text{Find} \quad & \boldsymbol{\nu} \in \mathbb{R}^n \\ \text{s.t.} \quad & \boldsymbol{A}_T^\top \boldsymbol{\nu} = \mathsf{sign}(\boldsymbol{x}_T) && (6.17) \\ & \|\boldsymbol{A}_{T^c}^\top \boldsymbol{\nu}\|_\infty < 1 && (6.18) \end{aligned}$$

Key idea 1: use iterative scheme to solve
minimize$_{\boldsymbol{\nu}}$ $\frac{1}{2}\|\boldsymbol{A}_T^\top \boldsymbol{\nu} - \mathsf{sign}(\boldsymbol{x}_T)\|^2$

for $t = 1, 2, \cdots$

$$\boldsymbol{\nu}^{(t)} = \boldsymbol{\nu}^{(t-1)} - \underbrace{\boldsymbol{A}_T \left(\boldsymbol{A}_T^\top \boldsymbol{\nu}^{(t-1)} - \mathrm{sign}(\boldsymbol{x}_T)\right)}_{\text{grad of } \frac{1}{2}\|\boldsymbol{A}_T^\top \boldsymbol{v} - \mathrm{sign}(\boldsymbol{x}_T)\|^2}$$

- Converges to a solution obeying the equality constraint; no inversion involved
- Issue: complicated dependency across iterations

Golfing scheme (Gross '11)

Key idea 2: sample splitting — use independent samples for each iteration to decouple statistical dependency

- Partition $\boldsymbol{A}$ into L row blocks $\underbrace{\boldsymbol{A}^{(1)} \in \mathbb{R}^{n_1 \times p}, \cdots, \boldsymbol{A}^{(L)} \in \mathbb{R}^{n_L \times p}}_{\text{independent}}$
- **for** $t = 1, 2, \cdots$ (stochastic gradient)

$$\boldsymbol{\nu}^{(t)} = \boldsymbol{\nu}^{(t-1)} - \underbrace{\mu_t \boldsymbol{A}_T^{(t)} \left(\boldsymbol{A}_T^{(t)\top} \boldsymbol{\nu}^{(t-1)} - \operatorname{sign}(\boldsymbol{x}_T) \right)}_{\in \mathbb{R}^{n_t} \text{ (but we need } \boldsymbol{\nu} \in \mathbb{R}^n)}$$

Golfing scheme (Gross '11)

Key idea 2: sample splitting — use independent samples for each iteration to decouple statistical dependency

- Partition $\boldsymbol{A}$ into L row blocks $\underbrace{\boldsymbol{A}^{(1)} \in \mathbb{R}^{n_1 \times p}, \cdots, \boldsymbol{A}^{(L)} \in \mathbb{R}^{n_L \times p}}_{\text{independent}}$

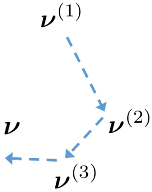
- **for** $t = 1, 2, \cdots$ (stochastic gradient)

$$\boldsymbol{\nu}^{(t)} = \boldsymbol{\nu}^{(t-1)} - \mu_t \tilde{\boldsymbol{A}}_T^{(t)} \left(\tilde{\boldsymbol{A}}_T^{(t)\top} \boldsymbol{\nu}^{(t-1)} - \text{sign}(\boldsymbol{x}_T) \right)$$

where $\tilde{\boldsymbol{A}}^{(t)} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{A}^{(t)} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{n \times p}$ is obtained by zero-padding

Golfing scheme (Gross '11)

$$\boldsymbol{\nu}^{(t)} = \boldsymbol{\nu}^{(t-1)} - \mu_t \tilde{\boldsymbol{A}}_T^{(t)} \Big(\underbrace{\tilde{\boldsymbol{A}}_T^{(t)\top} \boldsymbol{\nu}^{(t-1)} - \mathrm{sign}(\boldsymbol{x}_T)}_{\text{depends only on } \boldsymbol{A}^{(1)}, \cdots, \boldsymbol{A}^{(t-1)}} \Big)$$



- Statistical independence across iterations
  - By construction, $\boldsymbol{A}_T^{\top}\boldsymbol{\nu}^{(t-1)} \in \mathrm{range}(\boldsymbol{A}_T^{(i)\top}) \cap \cdots \cap \mathrm{range}(\boldsymbol{A}_T^{(L)\top})$
- Each iteration brings us closer to the target (like each golf shot brings us closer to the hole)

A general scheme for dual construction

$$
\begin{aligned}
\text{Find} \quad & \boldsymbol{\nu} \in \mathbb{R}^n \\
\text{s.t.} \quad & \boldsymbol{A}_T^\top \boldsymbol{\nu} = \mathsf{sign}(\boldsymbol{x}_T) & (6.19) \\
& \|\boldsymbol{A}_{T^c}^\top \boldsymbol{\nu}\|_\infty < 1 & (6.20)
\end{aligned}
$$

The golfing scheme doesn't yield an exact dual certificate, but an inexact one.

Theorem 6.11 (Inexact duality)

Let $T := \mathrm{supp}(\boldsymbol{x})$. Suppose $\|(\boldsymbol{A}_T^\top \boldsymbol{A}_T)^{-1}\| \leq 2$ and $\max_{i \in T_c} \|\boldsymbol{A}_T^\top \boldsymbol{a}_i\| \leq 1$. If

$$
\exists \boldsymbol{u} = \boldsymbol{A}^\top \boldsymbol{\nu} \text{ for some } \boldsymbol{\nu} \in \mathbb{R}^n \quad \text{s.t.} \quad \begin{cases} \|\boldsymbol{u}_T - \mathrm{sign}(\boldsymbol{x}_T)\| \leq 1/4, \\ \|\boldsymbol{u}_{T^c}\|_\infty \leq 1/2, \end{cases} ,
$$

then $\boldsymbol{x}$ is unique solution to ℓ_1 minimization.

Proof is similar to Theorem 6.9. The conditions in red is guaranteed with high probability via concentration inequalities.

Reference

[1] "*Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*," E. Candes, J. Romberg, and T. Tao, *IEEE Transactions on Information Theory*, 2006.

[2] "*Compressed sensing*," D. Donoho, *IEEE Transactions on Information Theory*, 2006

[3] "*Near-optimal signal recovery from random projections: Universal encoding strategies?*," E. Candes, and T. Tao, *IEEE Transactions on Information Theory*, 2006.

[4] "*Lecture notes, Advanced topics in signal processing (ECE 8201)*," Y. Chi, 2015.

[5] "*A mathematical introduction to compressive sensing*," S. Foucart and H. Rauhut, *Springer*, 2013.

[6] "*On sparse reconstruction from Fourier and Gaussian measurements*," M. Rudelson and R. Vershynin, *Communications on Pure and Applied Mathematics*, 2008.

Reference

[7] "*Decoding by linear programming*," E. Candes, and T. Tao, *IEEE Transactions on Information Theory*, 2006.

[8] "*The restricted isometry property and its implications for compressed sensing*," E. Candes, *Compte Rendus de l'Academie des Sciences*, 2008.

[9] "*Introduction to the non-asymptotic analysis of random matrices*," R. Vershynin, *Compressed Sensing: Theory and Applications*, 2010.

[10] "*Iterative hard thresholding for compressed sensing*," T. Blumensath, M. Davies, *Applied and computational harmonic analysis*, 2009.

[11] "*Recovering low-rank matrices from few coefficients in any basis*," D. Gross, *IEEE Transactions on Information Theory*, 2011.

[12] "*A probabilistic and RIPless theory of compressed sensing*," E. Candes and Y. Plan, *IEEE Transactions on Information Theory*, 2011.

Reference

[13] "*Statistical learning with sparsity: the Lasso and generalizations*," T. Hastie, R. Tibshirani, and M. Wainwright, 2015.

[14] "*Suprema of chaos processes and the restricted isometry property*," F. Krahmer, S. Mendelson, and H. Rauhut, *Communications on Pure and Applied Mathematics*, 2014.